THE HUMAN CELL ATLAS

Aviv Regev

Broad Institute of MIT and Harvard, Department of Biology, MIT, Howard Hughes Medical Institute

A fine resolution map of our basic functional units and their organization in tissues. To study human biology we must know our cells, and how they are organized and wired to make tissues. To comprehensively characterize and understand the foundational principles underling cellular organization we must to study both the component cells that that make up the tissue as well as tissue organization and function. The first aspect requires scale and careful experimental design and can be addressed by using current technologies to yield a **Human Cell Atlas**. The second aspect requires the development of new technologies, the scaling of emerging technologies and the generation of functional data that uses perturbations to decipher cellular circuits within and between cells that make up the tissue ecosystem. Such an effort towards a **functional tissue atlas** may be best served by dedicated pilot projects and technology development. We discuss these various aspects below.

1. THE HUMAN CELL ATLAS: A REFERENCE MAP OF OUR COMPONENT CELLS

The human body is made of more than 10^{13} cells and human cells are classified into distinct types based on their shape, location, function, and molecular profiles. Cells are a key intermediate in how genotypes manifest as phenotypes: genetic variants that contribute to disease typically manifest through their action in some cell types but not others. For example, a variant in a gene important for the function of dendritic cells and T-cells can increase the risk of an autoimmune disease such as multiple sclerosis, whereas another variant in a gene whose product functions primarily in skeletal muscle cells can lead to muscular dystrophy. Cells are also the starting points of most experimental work — be it isolating relevant cells from a human tissue, staining them in a section, growing them in culture, or targeting them within a genetically modified animal model.

Despite their importance, our characterization of cell types remains surprisingly limited. Much of our knowledge was acquired before the advent of molecular biology — some of it dates back decades, or even centuries — and it is often unknown if classifications by different criteria (e.g. shape, function and molecular characteristics) would even map neatly to each other. Genomics has offered the promise of a systematic approach. Yet thus far it has been applied largely in bulk to many cell types at once — masking many of the critical differences between cells — and in isolation from other valuable sources of data such as imaging.

Ideally, a first draft of a molecular atlas of cells in the human body would **(1)** catalog cell types and sub-types (*e.g.*, neurons, T-cells, *etc.*); (**2**) distinguish between cell states (*e.g.*, a naïve immune cell compared to the same immune cell type after encountering a bacterium); (**3**) relate cell types to their position; (**4**) capture the salient characteristics of cells during transitions such as differentiation or activation; (**5**) chart interactions between cells and, when possible, (**6**) trace the history of cells through a lineage.

Very recent advances in single-cell genomic analysis of cells and tissues have put within reach such a systematic, high-resolution effort to comprehensively characterize human cells. For example, massively parallel assays (*e.g.*, Drop-Seq (Macosko et al., 2015) and InDrop (Klein et al., 2015)) for RNA, CyTOF for proteins) can now process tens and hundreds of thousands of cells at very low cost. Rapidly emerging experimental and computational techniques that couple molecular profiling of RNA or proteins with spatial information from sub-tissue to sub-cellular resolution (*e.g.*, MERFISH (Chen et al., 2015b), FISSEQ (Lee et al., 2014), MIBI (Angelo et al., 2014), Tomo-Seq (Junker et al., 2014), Seurat (Satija et al., 2015) and more). Taken together, it is now possible to discover types, states, locations, transitions, interactions and lineage relations at an unprecedented resolution and scale.

Building a Human Cell Atlas nevertheless poses substantial challenges that can be overcome only in the context of a joint concerted effort. Because cells of the same type may reside in multiple tissues (depending on how "type" is defined), we must be able to compare across tissues in consistent ways. Because of the large number and diversity of tissue and cell types, we must bring together biomedical expertise in tissue and cell types, genomic and engineering expertise in efficient, high-quality, scalable data production, and computational expertise in data analysis, engineering and visualization. And we must harness the biomedical community's enthusiasm for such approaches in a coherent, coordinated way. Otherwise, we run the substantial risk that the resulting body of work will be inconsistent, incomparable, and erroneous, thus not only squandering a major opportunity, but even delaying and thwarting scientific progress.

A Cell Atlas would have immediate, tangible, and transformative benefits. **First**, it would provide a **reference map** for comparing related cells, identifying new cell types, helping interpret genetic variants, and identifying what distinguishes pathological cells from healthy ones. Knowing which genes are expressed in each tissue will further help to design better and safer therapeutics such as engineered CAR-Ts and address toxicity in drug development. **Second**, it would define sets of **markers and signatures**, facilitating the development of relevant reagents (*e.g.*, antibodies, probes) for molecular pathology, targeted cell sorting, and diverse additional assays. **Third**, because it would be derived directly from human tissue, it would provide a **direct view of human biology *in vivo***, removing the distorting aspects of cell culture and allowing us to develop better models for basic biology and drug discovery. **Fourth**, it would allow the effective **deconvolution of a massive body of legacy data** — from individual studies to catalogs such as TCGA — to resolve the true content of current profiles, greatly enhancing their impact. **Fifth**, it would help identify the **regulatory code** that controls cell differentiation, maintains cell state, and underlies cell–cell interactions, all key targets for fundamental understanding and therapeutic intervention. **Finally**, it would generate **"hardened," scaled, and broadly accepted methods** for sample preparation, lab protocols, informatics infrastructure, and data analysis.

2. HUMAN CELL ATLAS: DESIGN CONSIDERATIONS

Twenty-five years ago, scientists proposed the Human Genome Project as an audacious effort to systematically discover all of the cellular components encoded by the genome. The goal was daunting, but pioneering and visionary efforts defined a path forward -- making possible the generation of other comprehensive data resources, such as the Genotype-Tissue Expression Project (GTEx) (Consortium, 2013) and the Human Epigenomics Roadmap (Bernstein et al., 2010). A Human Cell Atlas would build on this legacy. It would (**1**) **systematically profile** at the single-cell level every tissue in the human body, including, whenever possible, during distinct phases of development and aging; (**2**) develop **analytical strategies** to define a reference map of cell types and states, their molecular signatures, lineage relations, and molecular regulatory underpinnings, and **data portals** that serve, visualize, and allow for exploration of the atlas; and (**3**) release extensive hardened **lab protocols and software** to allow studies across many domains.

Like the Human Genome Project, this work would be large but finite. It could only be completed successfully within the context of a unified project that engages a broad community of biologists, technologists, physicists, computational scientists and mathematicians (rather than as a collection of diverse projects supported by different institutes). The field of genomics has substantial experience in such projects, but a cell atlas is more diverse, complex and high-risk in technological and computational innovation than a typical genomics project.

In designing a Human Cell Atlas Project, we must address several important considerations:

(**i**) **Scale**.

An adult human has >10^{13} cells (exclusive of red blood cells), ~10^{11} of which are found in the brain. The range of the number of cell types depends on the level of resolution at which they are defined: an often-quoted estimate for the number of cell types is a few hundred, but the hematopoietic and immune system alone has >300 molecularly and functionally distinct sub-types in the mouse (Jojic et al., 2013)

, and the retina alone has ~100 defined neuron sub-types (Masland, 2012). While this is a daunting scale, it is chiefly a **sampling** problem and can therefore be addressed — we can sample orders of magnitude fewer cells and still recover fine distinctions with confidence, defined by solid statistical considerations. Mathematical theory also suggests that we can analyze far fewer than all transcripts or proteins and still recover correct results (Heimberg et al., 2016; Jaitin et al., 2014; Macosko et al., 2015).

We therefore propose that an Atlas rely on such a sampling framework to achieve a pre-defined level of sampling sensitivity, similar to strategies involved and refined in human genetics.

(**ii**) **Tissues and systems**

Samples. The core of a Human Cell Atlas should be profiling of normal, ideally fresh **human tissue.** Human normal tissue samples can be challenging for individual labs to acquire, but well-concerted efforts (*e.g.*, the GTEx project) can assemble excellent sample collections. Notably, unlike genetic studies, the Cell Atlas would not require a large number of individuals, making it more reasonable to procure fresh human samples. Emerging methods for single cell and single nucleus profiling allow analyzing fixed cells (after dissociation) or even lightly fixed tissue (without dissociation), opening the way to more flexible sample procurement. Nevertheless, one would ideally complement the effort with **model organisms** — primates, mice and others — both to obtain samples that are otherwise inaccessible, and to relate findings to experimentally tractable systems with extensive legacy knowledge. Finally, in relevant cases, ***ex vivo* or *in vitro*** experiments with primary cell culture or organoids could provide critical information to relate the most commonly used experimental systems to human *in vivo* biology.

We therefore propose to focus primarily on fresh human tissue, but complement these samples with model organism and in vitro *samples.*

(**iii**) **Staging**.

A Cell Atlas would be an endeavor of new scale and type. A **pilot phase** could be established quickly and serve to test alternative strategies and to evaluate the basic premises of the work. We envision such a pilot phase, with (**a**) a **sparser survey** of 10^5 cells from each of ~50 carefully chosen "**breadth tissues"** from human (and possibly mouse), and (**b**) a **deeper survey** in a few well-chosen complementary "**depth tissues"**, such as peripheral blood and bone marrow, gut, and liver. A **full-scale project** would analyze more cells per tissue and additional tissues from more individuals, expand work in model organisms, include stimulation where possible, deploy more measurement modalities, include more expansive funding for technology development, and possibly extend analysis to disease tissues within the domain of specific institutes.

Specifically, **depth tissues will be sampled based on their presumed complexity to a pre-defined level of completion**. That is, assuming a given number of cell types ***M*** in a tissue, the rarest type ***R*** being in proportion ***p*** (*e.g.*, *p* = 0.01%), we would measure a number ***C*** of cells needed to observe at least ***N*** cells of type ***R***. Tissues dominated by one or a few major subtypes may be experimentally depleted of the abundant type(s) to increase representation of rarer types. If markers are known for some populations (as in the case of blood and bone marrow), this information could be leveraged to distribute our sampling more efficiently, but we should always devote a substantial fraction to "negative" (and hence new) cells and/or to a completely unbiased sampling (no normalization of types by targeted enrichment or depletion). Depth tissue would also expected to include a larger number of auxiliary experiments and assays, such as follow up signature assays (CyTOF, *in situ*), and/or *ex vivo* assays (organoids, primary cultures).

Of the 30 "GTEx tissues" or comparable ("GTEx30"), excluding brain (see below), we propose to first tackle **5–10 depth tissues**. Depth tissues could be chosen by biological and technical consideration to span a range of organ systems, complexity of handling, and presumed diversity of cell types, in order to yield maximally useful information. Examples of possible depth tissues are blood, bone marrow, colon, liver, kidney, gut, lung and adipose. An effective sampling scheme would combine expert knowledge on types and available procedures for handling with statistical design criteria. We propose devote 50% of the cells (*e.g.*, 25M cells of a 50M cell project) to the depth tissues.

Conversely, **breadth tissues will be sampled at a fixed number of cells, to generate an initial draft, establish handling protocols, set up a biobank, and inform power analysis**. A similar number of samples would be procured and banked at the single cell level (*e.g.*, lysates or single-cell transcriptome attached to microparticles) for each breadth tissue as for a depth tissue. However, these samples would not proceed to sequencing at a similar scale. For example, we might expect to process ~20 M cells from ~20 breadth tissue types (1 M cells/breadth tissue). For most if not all breadth tissues, we expect minimal normalization (by enrichment or depletion), rather focusing on handling each sample as is. It is likely that for many tissue types, breadth sampling would provide substantial complexity.

We therefore propose a pilot phase, combining depth and breadth and laying the groundwork for a full-scale project.

(iv) Examples of potential uses.

A Human Cell Atlas would serve diverse goals, including a reference map for cell type and gene expression, a set of markers and signatures to define and manipulate them, a comparison into direct human biology *in vivo*, a means to deconvolute legacy bulk genomic data, and the ability to infer complex molecular circuits, within and between cells. The choice of depth and breadth tissue can impact these possibilities, as we highlight in the following illustrative use cases:

- ***Potential use: Understanding individual genetic variation*.** The initial goal of a Human Cell Atlas would be to generate a general map of cell types, but not yet to chart how cell types and profiles might vary by individual. In this, it is analogous to the role of the Human Genome Project in enabling projects that mapped human genetic variation (HapMap, 1000 Genomes Project). There are several ways in which a Cell Atlas could directly inform on the impact of genetic variation. For instance, genetic variation in gene expression in complex tissues may be due to cell-intrinsic effects (*e.g.*, on transcription) or to cell proportion effects (*e.g.*, increasing the ratio of one cell type over the other). It is difficult to distinguish between such effects using current data and approaches, but cell proportion effects could be identified by using the Cell Atlas to deconvolve existing GTEx and other complex profiles (*e.g.*, peripheral blood mononuclear cells). In addition, the pilot phase could include for one-depth tissue analysis across individuals after the initial characterization.
- ***Potential use: Understanding the brain*.** The brain clearly poses unique challenges due to its known complexity, inaccessibility, and difficulty to dissociate. In addition, the BRAIN initiative, GTEx, and the Common Fund Single Cell Program are all investing substantial efforts in single cell analysis of the brain. A Cell Atlas could build on and complement these efforts. While we cannot designate the brain as a depth tissue at the proposed bounds, we suggest devoting at least **5 M cells to the brain alone**, possibly using up to two brain parts as depth portions.
- ***Potential use: Understanding the hematopoietic and immune system*.** We propose two immune tissues (blood and bone marrow) as depth tissues. Both tissues are uniquely accessible and rich in knowledge and tools. In addition, blood is regularly assessed in clinical contexts, and many of the breadth tissues will likely yield tissue resident immune cells; these considerations suggest that blood and bone marrow should be given special attention. These could be complemented with key epithelial and other barrier tissues that are enriched in immune cells, or in depth analysis of CD45+ cells across many tissues.
- ***Potential use: Understanding epithelial tissues.*** We propose two epithelial tissues (*e.g.*, intestinal crypt and lung) as depth tissues. These are both mucosa-lined luminal organs with a shared embryological origin, and exposure to microbiome. In both tissues there are extensive interactions between epithelial cells and epithelial cell and immune cells. Moreover, there seems to be cross talk between these organs. Finally, the intestinal crypt is uniquely accessible and rich in knowledge and tools. These considerations suggest that should be given special attention.

Potential use: ex vivo analysis. We do not exclude the possibility of ***ex vivo* stimulated or perturbed cells** to capture transient states using **organoids** or other more experimental approaches. Since these are expected to be far less complex in cell type, and since we will not include genetic perturbation in this Atlas, we do not expect these to substantially impact the numerical scope of the project substantially.**(v) Technical capabilities.**

Measurements. Three main measurement strategies would provide critical data: (**1**) **Single cell RNA-Seq (scRNA-seq)** provides genomic profiles of individual cells, and massively parallel methods such as Drop-Seq, inDrop (Klein et al., 2015) and their commercial variants, make it compatible with analyzing tens of thousands of cells in one low-cost, high-throughput multiplex process. While scRNA-seq is ideally applied to freshly dissociated tissue, there are emerging strategies for work with fixed samples (Habib et al., 2016; Thomsen et al., 2016). (**2**) **Mass cytometry (CyTOF)** and related methods allow multiplexed measurement of proteins based on antibodies barcoded with heavy metals (Bendall et al., 2014; Gut et al., 2015). These measure only pre-defined signatures, but can process many millions of cells for a very low cost per cell. (**3**) ***In situ* techniques**, relying on multiplexed RNA-FISH or antibody staining in fixed cells and tissues, require a pre-defined signature and are more limited in the number of cells assayed. However, they provide spatial resolution, even at sub-cellular levels (Chen et al., 2015b; Lee et al., 2014).

*We therefore propose scRNA-Seq as the workhorse method for initial characterization, and subsequent definition of signatures for studies that require additional scale (mass cytometry) or spatial resolution (*in situ*). We discuss the specific challenge of spatial analysis below.*

(**iv**) **Organization**.

A Cell Atlas Project would need to draw on and balance diverse expertise that balances the need for domain expertise and development of new technologies (more so than in past genomics projects) with the need for data collection, management and analysis that is comparable across systems. Within a consortium, to be defined through a community process, there would need to be working groups for human (adult) samples (surgery and pathology), human (fetal) samples, model organisms (*e.g.*, mouse, primates), and cell culture, as well as for technology development. We should also expect a central data acquisition, management and release effort, along with multiple analytics efforts.

We therefore propose a concerted effort, to be defined through a community process. The consortium would have relevant working groups, shared data acquisition, coordination, management and serving, and a broad diversity of analytical tool development.

(vi) Data needs.

For the generated data to truly form a Human Cell Atlas would require novel analysis and visualization capabilities that would allow data-driven definitions of all aspects of a cell's identity as well as the ability to navigate, query and visually inspect this information. We envision that the underlying data would usher a flurry of new analytical methods, which will embrace and leverage the unique structure, noise properties and massive scale of this new data, and thus extending far beyond the realm of adoption and adaptation of tools developed for other (non biological) purposes. Similarly, there will be a substantial need for new modalities for data visualization and exploration. Such efforts are best served by maximally open access to the collected data, without embargos and access restrictions.

As new methods are developed at better scale, spatial resolution and function, we also need to facilitate the dissemination and sharing of data collected by any initiative with the scientific community in dedicated portals. Such portals can house all the data, allow visualization and sharing and include analytic tools. A Human Cell Atlas Portal would allow for upload and download of datasets and metadata, as well as intuitive data exploration and visualization. Data would be available on the Cloud, and compute could be performed on the cloud as well, using on-the-fly analysis (such as those currently being developed for the NCI's Cloud Pilot).

We therefore propose that a Human Cell Atlas project will include open-access dissemination of all data, not only supporting directly computational method development, but allowing any member of the community to develop methods for data analysis, computation, and visualization. We also propose central investment in data sharing, dissemination and portals, and Cloud-based computing.

(vii) Existing and emerging efforts.

There are emerging efforts in the community, which could help seed a Human Cell Atlas. Among these:

- Our group is involved, through the Immgen Consortium and philanthropic support, in an effort to chart **pilot immune cell atlases,** from both immune and non-immune tissue for the diversity of immune cells. Such an effort provides expertise and infrastructure at collecting, analyzing, and serving millions of single cell profiles, and at procuring and processing diverse samples. It also provides valuable strategies on the design of sampling experiments. From our collaborator, Dr. Christoph Benoist, and others, we are aware that the Allen Institute and the NIAID both expressed interest in expanding such efforts beyond a pilot scale.
- Our group serves as a hub for an effort to **pilot cancer cell atlas:** single cell analysis across over 20 different tumor types, including epithelial tumors, brain tumors and hematologic malignancies. This has built a network of collaborations with hospitals, and developed a pipeline to process diverse fresh human samples, including resections, core biopsies, fine needle aspirates, and bodily fluids, that can be applicable to normal tissue as well. We are increasingly aiming to conduct surveys of matching normal tissue, whenever possible, at pilot scale.

- The **BRAIN Initiative** supports a series of projects that focus on single cell analysis of the mammalian brain. In our case, we have partnered with Dr. Joshua Sanes, Alex Schier and additional collaborators to perform a census of the mouse retina. Our groups are charting other areas of the mouse brain and some human samples. This work has provided protocols to handle some of the toughest tissue types, such as our single nucleus protocols, and to profile cells at massive scale (with Drop-Seq).
- There is an emerging effort to coordinate **a Human Cell Atlas project across the international community**, as many groups, including ours, work to dissect diverse tissue types. A small steering group, with leadership from the Broad Institute, Sanger Institute and the Wellcome Trust is organizing an international meeting for October, aiming to invite scientists from all relevant areas, along with funders, philanthropist, and representatives from major technology and pharmaceutical companies to plan and lay the foundation for the generation of a human cell atlas.

3. CHARACTERIZING TISSUE ORGANIZATION: SPATIALLY RESOLVED GENOMICS

Physiological processes in health and disease take place not only in complex cells but also in complex tissues, where interacting cells and molecules are organized in space and time. In particular, spatial analysis of tissue sections is a cornerstone of pathology in the clinical practice. The plummeting cost of DNA sequencing has driven development of a wide spectrum of technologies aimed at encoding into DNA diverse types of biological information, from RNA levels to physical interactions between proteins to the 3D organization of the chromosome, and then reading this information back out by sequencing. However, these technologies – even when applied at single cell resolution – generally dissociate and/or homogenize specimens prior to analysis and thus discard crucial information about how molecules and cells fit together and interact in space.

This has led to a major gap in our ability to leverage genomics in basic biology and the clinic to address critical questions in human health and disease, and to impact patient care. There is thus an **enormous need** for new methods for analyzing **spatial distributions** of any **genomic diversity** in biological tissue.

We envision two key parallel paths for spatially resolved analyses:

- **High resolution secondary analysis.** Many of the emerging spatial techniques rely on analysis of pre-defined gene signatures. As a result, they are the ideal complement to profiling of single cells from dissociated tissue. In this design, a large number of cells is first analyzed from a tissue, and then used to define the most informative signatures for spatial analysis. At a second step, tissue section or whole mount samples are analyzed by the signature with spatial approachaes. Computational methods can then impute the behavior of the genes that were not measured, especially in many types of healthy tissue with canonical structures.
- **Technology development**. There is also an important need for further development of methods that are highly scaled, broadly deployed and ideally, collect genome wide profiles. While a general solution is not yet fully realized, this is an area of intense work for many groups, with several emerging key strategies, which we briefly review below. These strategies could form the basis of a significant technology development effort that would complement a Human Cell Atlas and provide critical information on tissue organization.

(i). Experimental approaches

Experimental approaches rely on **direct coding** of the position of a molecule: obtain a barcode (molecular or physical) such that a cell and/or molecule can be registered to space, and is associated with any ‘omics data derived from that cell/location.

Multiplex RNA fluorescence in situ hybridization (RNA FISH). A set of recently developed methods builds on RNA-FISH with sequential hybridizations and error correction encodings to hybridize up to thousands of different transcripts in multiplex, most dramatically in the recently published MERFISH method (Xiaowei Zhuang lab) (Chen et al., 2015b). Published data indicated good spatial resolution (with higher resolution correlated to increasing imaging time) and good quantitative accuracy compared to single cell RNA-seq.

Improvements in Expansion Microscopy (Ed Boyden's lab) (Chen et al., 2015a) can in principle be combined with these approaches.

In situ sequencing. Two recent studies performed RNA-seq *in situ* in preserved tissue sections and in cells (George Church and Mats Nilsson groups). The published FISSEQ proof-of-concept (Lee et al., 2014), while very impressive, suffered from very low complexity, and raised concerns about substantial skews in representation, possibly due to issues in imaging or accessibility of RNA to sequencing. Updates in chemistry and sample preparation (especially with Expansion Microscopy) are expected to improve its performance.

In situ tagging prior to scRNA-Seq*.* An alternative to sequencing *in situ*, is to *tag* RNA transcripts *in situ*, with spatially resolved barcodes, followed by isolation of cells or RNA, such that following sequencing, one can map back the transcripts/cells to the image. In one such method a tissue section is hybridized to an oligonucleotide array, providing XY coordinates to the isolated RNA (Jonas Frisen lab). Other labs are developing approaches that would rely on optical tagging of cells.

In situ multiplex protein localization. Multiplex protein detection using mass cytometry with heavy metal barcoded antibodies (detected on a Time Of Flight (TOF) mass spectrometer) can now be coupled with either laser ablation (Bernd Bodenmiller lab) (Bodenmiller, 2016) or ion beam ablation (Gary Nolan, Michael Angelo and Sean Bendall's labs) (Angelo et al., 2014) of tissue sections. The laser (lower resolution) or ion beam (higher resolution) raster the tissue section – which was pre stained with the barcoded antibodies – pixel by pixel, followed by detection using TOF. This allows reconstruction of an "image" with the quantities of the proteins in each pixel. Both techniques are expected to be commercially available. The methods should be compatible in principle with nucleic acid detection as well.

(ii). Computational inference approaches

Computational strategies rely on **indirect inference** of spatial location: they combine spatially resolved *in situ* information at least at the molecular signature level, as well as single cell data from dissociated tissue, and then use the overlapping information (*e.g.*, RNA signature) to map the cells to their location using landmarks. This requires the two measured samples to be "the same", either because tissue is canonically structured, or because the niches change relatively continuously, and the two samples are physically adjacent.

Computational inference of spatial positions in canonical tissue*.* In the past year, several computational methods have been proposed for the inference of spatial positions of transcripts or cells by combining experimental measurements with computational algorithms. Tomo-Seq (van Oudenaarden's lab) (Junker et al., 2014) reconstructs the spatial position of RNA in tissue from RNA-Seq of cryosections in each relevant dimension. Seurat (Regev lab) (Satija et al., 2015), and a similar method from Marioni lab (Achim et al., 2015), both combine spatial reference maps of a small number of transcripts collected by individual RNA *in situ* hybridization (ISH or FISH) with single cell RNA-Seq of dissociated cells to infer the position of each cell and the spatial distribution of each transcript. All three approaches rely on the *canonical structure* of the tissue: the abilty to obtain the "same" tissue once for generating reference maps and independently for scRNA-Seq. Our lab has now applied this approach in mammalian tissue (*e.g.*, mouse hippocampus).

Computational inference of spatial positions in idiosyncratic tissue*.* While many normal tissues have this canonical behavior, cancer samples are by nature idiosyncratic. In this case, we may assume that the tissue is continuously organized and relying on matched alternating samples (taken for in situ measurement and single cell 'OMICS analysis, respectively). It is possible that using these matched measurements as a training set, there are further high-order features that are common across many samples (e.g., an invasive edge in a tumor). Assigning cells and transcripts to these higher order features may carry the most important information.

(iii). Considerations for comparing and assessing methods

The methods discussed above vary based on several key parameters:

- **Specialized equipment**. FISSEQ and MIBI currently rely on highly specialized equipment not yet broadly available to other labs. This will likely improve with commercialization, but could still pose a cost barrier. Resolution for such methods may improve when combined with new preparation approaches, such as Expansion Microscopy.

- **Throughput**. Some of the imaging-based strategies, such as MIBI, FISSEQ and MERFISH, require substantial imaging/processing time for a relatively small field of view. This is likely to improve substantially in new versions.
- **Ability to handle idiosyncratic samples.** Currently published computational methods, such as Seurat, rely on reference maps, and thus are not applicable to wholly idiosyncratic samples (*e.g.*, tumors). New computational developments (above) should alleviate this limitation.
- **Complexity.** Signature based methods (MERFISH, MIBI) require a pre-defined set of target transcripts or proteins for analysis. FISSEQ, as published, had limited complexity and non-trivial biases, although those are expected to improve.
- **Spatial and single cell resolution.** Approaches that are based solely on an imaging readout, may have resolution limitations given the associated microscopy and sample preparation. Many of these methods are not single cell assays, and image processing is required for single cell assignment. In situ tagging in live samples may provide both spatial and single cell resolution.

*Overall, there is intensive development of methods for spatially-resolved genomics and proteomics. **Computationally**, existing inference methods likely suffice for many canonical tissues, and ongoing work will tackle idiosyncratic tissue (e.g., tumors). **Experimentally**, in situ sequencing is one strategy, but massive-multiplex FISH, RNA in situ tagging and other emerging techniques may offer other critical opportunities. Current methods can thus be combined as secondary assays leveraging signatures from single cell profiles; in addition, there should be investment in technology development and deployment*.

4. FUNCTIONAL TISSUE ATLAS: THE NEED FOR PERURBAIONS

A Human Cell Atlas would focus on characterization of cellular identity, with spatial information in situ an important component. While this data likely will allow inference of important functional information, understanding tissue function – a question raised in the RFI – required requires perturbations, such that causality can be established. Perturbations can rely on both natural variaion, including genetic differences between individuals, or on engineered perturbations at the gene, cell and interaction level.

As a result a *functional* tissue atlas – as opposed to a cell atlas -- would require not just profiling at large scale and high spatial resolution, as discussed above, but also systematic perturbations at massive scale and with sophisticated readouts, applied to faithful biological samples, whenever possible. While we firmly believe that such techniques will be emerging rapidly in the near future, this is a distinct endeavor, and would likely be best suited to more focused, system-specific efforts at least for the time being.

5. WHY COMMON FUND? THE BENEFITS OF A HUMAN CELL ATLAS

A Human Cell Atlas would be akin to having the “zip code” of each cell type. It would provide foundational biological knowledge on the composition of multicellular organisms — enabling us to understand how cell types weave together in three dimensions to form tissues, how the map is generated to connect all of the body's systems, and how changes in the map underlie health and disease. The atlas would also enable us to develop effective diagnostics and treatments. Specifically, it would:

- **Serve as a reference map for discovery and characterization of cell types and states.** An Atlas would identify new cell types, states, and transitions, as well as their molecular characteristics, thereby defining the functions of known and novel cell types. It would also help compare cells of similar types between contexts, such as tissues, health and disease state, different individuals, and between human and model organisms. Cell type expression profiles would also help determine the cell types in which disease-associated genes are acting. It can also help determine the cell types in which specific genes are expressed thus providing better guidance in to the design of therapeutics (e.g. CAR-T) and better prediction of drug toxicities.

- **Define markers and signatures.** To study the functions and pathologies of cells, we must be able to isolate the cells from samples and recognize them in tissue sections. A Human Cell Atlas would define sets of markers and signatures, allowing us to develop relevant reagents, such as antibodies and probes for molecular pathology and cell sorting. In addition, such reagents can be used in multiplex in single-cell assays at the signature scale (multiplex FISH, CyTOF, multiplex qPCR), increasing by orders of magnitude the number of cells and samples that can be further analyzed.

- **Provide a direct view of human biology *in vivo***. Much of our understanding of cells and their roles still relies on studies in cell culture, far removed from *in vivo* human physiology. A Human Cell Atlas would remove many of the distorting aspects of cell culture, by analyzing fresh human tissues, processed immediately, at high resolution. It would therefore allow us to compare our current models to *in vivo* biology and to develop more faithful models for basic biology studies and drug discovery.
- **Allow effective deconvolution of a massive body of legacy data.** In the past decade, both large-scale efforts (such as TCGA and the Epigenomics Roadmap) and many investigator-initiated studies have collected a vast body of genomic profiles of complex tissues, from PBMCs to tumors. A Human Cell Atlas with appropriate computational analysis could serve to deconvolve these complex tissues to their constituent cellular contents, thus vastly increasing their resolution and truly maximizing their value. Notably, because tissue cells typically assume different profiles than cells cultured *in vitro*, only a Human Cell Atlas would allow such reliable deconvolution. A Human Cell Atlas project should first establish the validity of such convolution by comparing the outcome for tests sets *vs*. ground truth from tissue sections or FACS.
- **Help uncover the regulatory code that controls cell differentiation, state, and interactions.** The regulatory circuits that control cell differentiation, states, and interactions are a primary area of interest for both basic research and therapeutic intervention. Despite intensive efforts for the past two decades, resolving regulatory circuits, especially in tissue *in vivo*, remains a formidable challenge. The variation between single cells of the “same” and “different” types and states provides a unique signal with which to recover regulatory relations.
- **Generate hardened, scaled, and broadly applicable experimental and computational methods**. As with the human genome, the diversity of biology — across human individuals, developmental and pathological conditions, genetic, physiological and pharmacological perturbations, as well as additional species — naturally extends well beyond a pilot or even a full-scale Human Cell Atlas. We therefore expect methods that arise from a Human Cell Atlas Project to be useful in many other contexts across a broad range of areas, for any tissue, condition or organism of interest. We also anticipate that after 5 years, empowered by the data and technical accomplishments of the Atlas, projects funded by individual centers and institutes would be able to tackle further “deep dives” in specific biological and disease areas such as cancer.

In sum:

A Cell Atlas would be transformative. Every aspect of biomedical research relies on knowing cell types and their distinct molecular markers. Despite this fundamental fact, many of our classifications still date to before the era of molecular biology, and new types and markers are often identified in haphazard, sometimes inconsistent ways that confound cause and correlation. A coherent, comprehensive Atlas would (a) result in a core set of cell-type definitions that will be actionable experimentally, (b) allow us to deconvolve most previously collected bulk sample measurements, and (c) generate experimental and computational tools to enable high-resolution research from the bench to the clinic.

A Cell Atlas would be synergistic. Several existing NIH initiatives relate to the question of cell type. For example, the Immunological Genome Consortium aims to identify molecular profiles and underlying mechanisms of immune cell types in mouse (though the cell types are pre-defined). One of the key RFAs in the BRAIN initiative focuses on methods for cell type characterization. GTEx has characterized gene expression profiles across many different tissues. ENCODE is primarily pursued in well-defined cell types, and has raised the need for clear definition of their underlying heterogeneity. While each of these efforts complements a Cell Atlas, it neither generates such an atlas nor could supplant it. The need to systematize and harden methods requires an effort well beyond that of these individual projects. Furthermore, a Cell Atlas Project would provide tools, methods, standards, and data to support these specialized efforts and many others.

A Cell Atlas would be cross-cutting. A Human Cell Atlas would apply to all tissues, diseases, and biological systems. Once the Atlas is established, it could serve as a benchmark for assessing any disease-specific sample or state. Notably, cell types cannot be neatly partitioned by ICs, disease, or biological area. Key loci of disease involve cells from multiple IC areas (for example, the gut contains epithelial cells, diverse immune cells, neurons, and the microbiome), such that the "same" cell type is present in distinct tissues and assumes different states that transition across our classifications (*e.g.*, during development).

A Cell Atlas would be catalytic. Within five years, a scaled Project (*e.g.*, 50M cells) could generate the draft Atlas providing profiles, signatures, and classifications that will allow finer analysis in disease pathology, rare cell types, and deconvolution of bulk samples. The latter would greatly increase the value of a precious body of legacy data from the past two decades of genomics. The project would have also generated methods for sample preparation, lab protocols, informatics infrastructure, and analysis that are "hardened," scaled, and broadly accepted, such that further work could deploy them in ongoing research.

REFERENCES

Achim, K., Pettit, J.B., Saraiva, L.R., Gavriouchkina, D., Larsson, T., Arendt, D., and Marioni, J.C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nature biotechnology *33*, 503-509.

Angelo, M., Bendall, S.C., Finck, R., Hale, M.B., Hitzman, C., Borowsky, A.D., Levenson, R.M., Lowe, J.B., Liu, S.D., Zhao, S.*, et al.* (2014). Multiplexed ion beam imaging of human breast tumors. Nat Med *20*, 436-442.

Bendall, S.C., Davis, K.L., Amir el, A.D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell *157*, 714-725.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R.*, et al.* (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nature biotechnology *28*, 1045-1048.

Bodenmiller, B. (2016). Multiplexed Epitope-Based Tissue Imaging for Discovery and Healthcare Applications. Cell Syst *2*, 225-238.

Chen, F., Tillberg, P.W., and Boyden, E.S. (2015a). Optical imaging. Expansion microscopy. Science *347*, 543-548.

Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015b). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science *348*, aaa6090.

Consortium, G.T. (2013). The Genotype-Tissue Expression (GTEx) project. Nature genetics *45*, 580-585.

Gut, G., Tadmor, M.D., Pe'er, D., Pelkmans, L., and Liberali, P. (2015). Trajectories of cell-cycle progression from fixed cell populations. Nature methods *12*, 951-954.

Habib, N., Li, Y., Heidenreich, M., Swiech, L., Trombetta, J.J., Zhang, F., and Regev, A. (2016). Div-Seq: A single nucleus RNA-Seq method reveals dynamics of rare adult newborn neurons in the CNS. bioRxiv.

Heimberg, G., Bhatnagar, R., El-Samad, H., and Thomson, M. (2016). Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. Cell Syst *2*, 239-250.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., *et al.* (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776-779.

Jojic, V., Shay, T., Sylvia, K., Zuk, O., Sun, X., Kang, J., Regev, A., Koller, D., Immunological Genome Project, C., Best, A.J., *et al.* (2013). Identification of transcriptional regulators in the mouse immune system. Nat Immunol *14*, 633-643.

Junker, J.P., Noel, E.S., Guryev, V., Peterson, K.A., Shah, G., Huisken, J., McMahon, A.P., Berezikov, E., Bakkers, J., and van Oudenaarden, A. (2014). Genome-wide RNA Tomography in the zebrafish embryo. Cell *159*, 662-675.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187-1201.

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R., Jeanty, S.S., Li, C., Amamoto, R., *et al.* (2014). Highly multiplexed subcellular RNA sequencing in situ. Science *343*, 1360-1363.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., *et al.* (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202-1214.

Masland, R.H. (2012). The neuronal organization of the retina. Neuron *76*, 266-280.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nature biotechnology *33*, 495-502.

Thomsen, E.R., Mich, J.K., Yao, Z., Hodge, R.D., Doyle, A.M., Jang, S., Shehata, S.I., Nelson, A.M., Shapovalova, N.V., Levi, B.P., *et al.* (2016). Fixed single-cell transcriptomic characterization of human radial glial diversity. Nature methods *13*, 87-93.