

*The National
Academies of*

SCIENCES
ENGINEERING
MEDICINE



Forum on
NEUROSCIENCE and
NERVOUS SYSTEM DISORDERS

Neuroscience Data in the Cloud—A Workshop

Session II Breakout

Data Management

Moderator: MICHAEL HAWRYLYCZ, Allen Institute for Brain Science

Rapporteur: MICHAEL HUERTA, National Library of Medicine Discussants:

DANIEL MARCUS, Washington University School of Medicine

RACHEL RAMONI, Department of Veterans Affairs

JANAINA MOURAO-MIRANDA, University College London (*invited*)



Use Cases: Single cell technologies have spawned enormous efforts to map cellular diversity and enormous amounts of complex data

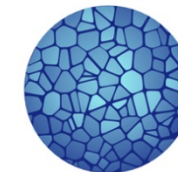
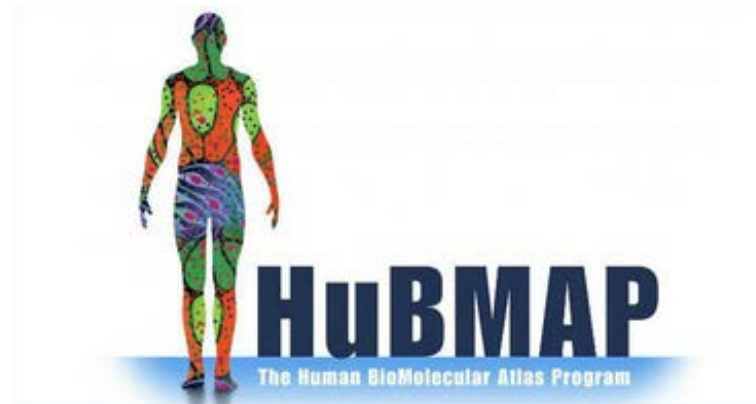
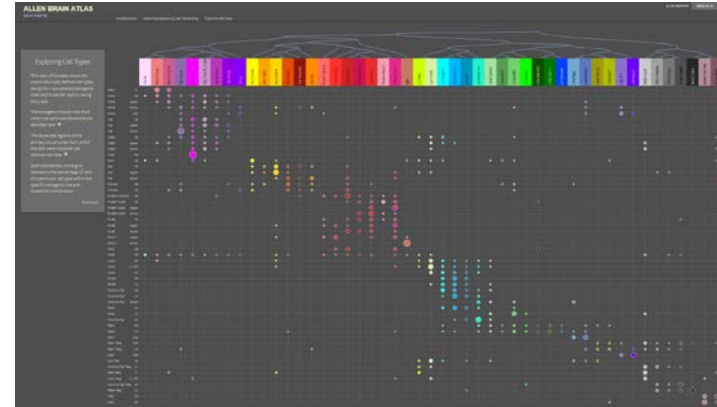


Neuron
Perspective

The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas

Joseph R. Ecker,¹ Daniel H. Geschwind,² Arnold R. Kriegstein,³ John Ngai,^{4,*} Pavel Osten,⁵ Damon Polloudakis,² Aviv Regev,⁶ Nenad Sestan,⁷ Ian R. Wickersham,⁸ and Hongkui Zeng⁹
¹Genomic Analysis Laboratory and Howard Hughes Medical Institute, Salk Institute for Biological Studies, La Jolla, CA 92037, USA
²Program in Neurogenetics, Departments of Neurology and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
³Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, Department of Neurology, University of California, San Francisco, San Francisco, CA 94143, USA
⁴Department of Molecular and Cell Biology, Helen Wills Neuroscience Institute, QB3 Functional Genomics Laboratory, University of California, Berkeley, Berkeley, CA 94720, USA
⁵Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
⁶Klamman Cell Observatory, Broad Institute of MIT and Harvard, Department of Biology, Koch Institute of Integrative Cancer Research, and Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
⁷Departments of Neuroscience, Genetics, Psychiatry and Comparative Medicine, Program in Cellular Neuroscience, Neurodegeneration and Repair, Yale Child Study Center, Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT 06510, USA
⁸McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
⁹Allen Institute for Brain Science, Seattle, WA 98109, USA
*Correspondence: jngai@berkeley.edu
<https://doi.org/10.1016/j.neuron.2017.10.007>

Allen Cell Types Database



HUMAN
CELL
ATLAS



bioRxiv
beta
THE PREPRINT SERVER FOR BIOLOGY

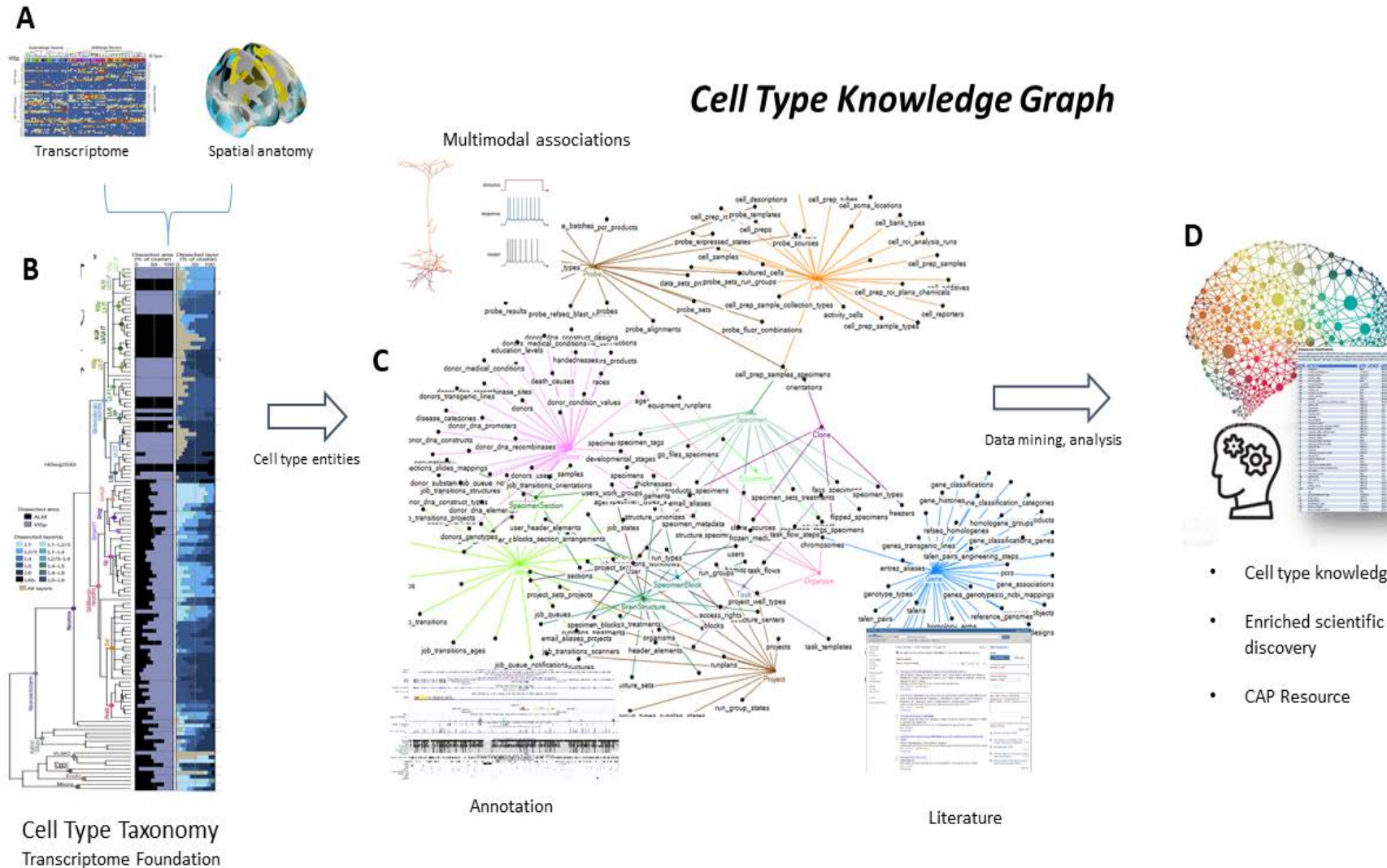
New Results

The Human Cell Atlas

Aviv Regev, Sarah Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Ellis, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Gottgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemmer, Seung K. Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Joakim Lundberg, Partha Majumder, John Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Philipakis, Chris P. Ponting, Stephen R. Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua R. Sanes, Rahul Satija, Tom Shumacher, Alex K. Shalek, Ehud Shapiro, Padmanee Sharma, Jay Shin, Oliver Stegle, Michael Stratton, Michael J. T. Stubbington, Alexander van Oudenaarden, Allon Wagner, Fiona M. Watt, Jonathan S. Weissman, Barbara Wold, Ramnik J. Xavier, Nir Yosef, Human Cell Atlas Meeting Participants
doi: <https://doi.org/10.1101/121202>



Use Cases: Knowledge Graphs for Data Management and Inference



Discussion Questions

- What are our expectations for cloud based data in neuroscience?
- What particular aspects of cloud data management and best practices are specific to neuroscience?
- What use cases will drive cloud data management for neuroscientists?
- Which aspects of human data access and management are particularly challenging?



Characteristics of the cloud based data management

- Compute power is elastic, but only if workload is parallelizable
- Data is stored at an untrusted host
- Data is replicated, often across large geographic distances



Desired features of cloud based computing environments

- Efficiency
- Fault tolerance
- Ability to run in a heterogenous environment
- Ability to run on encrypted data
 - Human data privacy concerns
- Interfaces well with other applications and common programming environments



Data Management Tools and Architecture

- **(Apache) Hadoop**
 - An open source platform providing highly reliable, scalable, distributed processing of large data sets using simple programming models.
- **(Apache) MapReduce**

A programming paradigm that allows for massive scalability of unstructured data across hundreds or thousands of commodity servers in an Apache Hadoop cluster.
- Many commercial systems Google, Amazon, Adobe, IBM,...
- Shared-nothing parallel databases

