THE UNIVERSITY OF TOKYO

DOCTORAL THESIS

博士論文

CUBIC-Cloud: An Integrative Computational Framework Towards Community-Driven Whole-Mouse-Brain Mapping

CUBIC-Cloud: 分散型マウス全脳マッピングのためのクラウド解析システム

Author:
Tomoyuki Mano
真野智之

Supervisor:
Dr. Hiroki R. Ueda
上田泰己

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Graduate School of Information Science and Technology, Department of Information Physics and Computing

March 1, 2021

Declaration of Authorship

I, Tomoyuki Mano, declare that this thesis titled, "CUBIC-Cloud: An Integrative Computational Framework Towards Community-Driven Whole-Mouse-Brain Mapping" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"By the means of Telescopes, there is nothing so far distant but may be represented to our view; and by the help of Microscopes, there is nothing so small, as to escape our inquiry; hence there is a new visible World discovered to the understanding. By this means the Heavens are open'd."

Robert Hooke, *Micrographia* (1665)

THE UNIVERSITY OF TOKYO

Abstract

Graduate School of Information Science and Technology, Department of Information Physics and Computing

Doctor of Philosophy

CUBIC-Cloud: An Integrative Computational Framework Towards Community-Driven Whole-Mouse-Brain Mapping

by Tomoyuki Mano

Recent advancements in tissue clearing technologies have offered unparalleled opportunities for researchers to explore the whole mouse brain at cellular resolution. With the expansion of this experimental technique, however, a scalable computational framework is demanded to efficiently analyze and integrate whole-brain mapping datasets collected by the research community. To that end, here I present CUBIC-Cloud, a cloud-based framework to quantify, visualize and share whole mouse brain data.

In chapter 1, I will review the previous whole mouse brain mapping projects, and explain how tissue clearing and light-sheet microscopy imaging method is bringing a new breakthrough in this landscape. Based on these observations, I will outline the software challenges that needs to be addressed to achieve more scalable and efficient whole brain mapping. Chapter 2 describes the experimental methods and software implementation details. In chapter 3, I will explain the CUBIC-Cloud framework and describe the front-facing functionalities to allow researchers to upload, analyze and publish the whole brain mapping data. I also show that the serverless architecture used in CUBIC-Cloud allows the system to dynamically scale its computational power depending on the load by the user. In chapter 4, using CUBIC-Cloud framework, I will present some novel whole mouse brain mapping results to demonstrate the usability of the proposed framework. First, I investigated the brain-wide distribution of various cell types, including PV, SST, ChAT, Th and Iba1 expressing cells. Second, I reconstructed neuronal activity profile under pharmacological perturbation using c-Fos immunostaining. Third, a brain-wide connectivity mapping by pseudo-typed Rabies virus will be demonstrated. Together, CUBIC-Cloud provides an integrative platform to advance scalable and collaborative whole-brain mapping.

Acknowledgements

I wish to thank my supervisor, Dr. Hiroki Ueda, for his help and advice during my PhD research. His passion and tireless efforts have provided a resourceful, unique and synergistic research environment to pursue my doctoral research. I feel greatly honored that I was able to be a part of this amazing research team.

I am greatly thankful to Dr. Shoi Shi, who has provided enormous advice on my research. I would also like to thank the members of Ueda laboratory, Dr. Etsuo Susaki, Kazuhiro Kon, Dr. Masafumi Kuroda, Chika Shimizu, Rina Tanaka, Dr. Tatsuya Murakami, Dr. Takeyuki Miyawaki and Hiroaki Ono, who I worked together in the tissue clearing projects. I am grateful to Dr. Ken Murata, Dr. Kazunari Miyamich and Dr. Kazushige Touhara for the collaboration on the rabies virus experiments. I also thank graduate students from IST, Shion Honda, Tomoyuki Asanuma and Machiko Katori, for their friendship and fruitful research discussions.

My doctoral study has been financially supported by Japanese Society for the Promotion of Science (JSPS) through Research Fellowship for Young Scientists (DC2) and by ANRI scholarship.

Lastly, I would like to thank my family for their love and support throughout my graduate degree. My deepest gratitude goes to my mother and father, who have always encouraged me to pursue my interest in science.

Contents

List of Figures

List of Tables

List of Abbreviations

AAV(s)	Adeno-associated viral vector(s)
API(s)	Application programming interface(s)
ASD	Autism spectrum disorder
CNN	Convolutional neural networks
ChAT	Choline acetyltransferase
CNS	Central nervous system
CT	Computed tomography
DMN	Default mode network
GABA	Gamma-aminobutyric acid
GPU	Graphic processing unit
FOV	Field of view
FP	Fluorescent protein
Iba1	Ionized calcium-binding adapter molecule 1
IEG(s)	Immediate early gene(s)
LSFM	Light-sheet fluorescence microscope
LPS	Lipopolysaccharides
MRI	Magnetic resonance imaging
PFA	Paraformaldehyde
PBS	Phosphate-buffered saline
PV	Parvalbumin
RI	Refractive index
RV	Rabies virus
scRNA-seq	Single-cell RNA sequencing
SST	Somatostatin
TH	Tyrosine hydroxylase
VIP	Vasoactive intestinal peptide

Chapter 1

Introduction

The researches on anatomical understanding of the brain blossomed in the early 20th century by the pioneering work by Cajal and Golgi. Since then, the neuroscientists have explored the microscopic structures of various kinds of neurons of diverse organisms. Combined with single-cell recording of the electrophysiological properties and molecular identification of the ion channels and pumps, our understanding of a single neuron, a computational unit of the brain, has been vastly advanced. However, it is almost certainly true that a single cell alone cannot produce "intelligence" as we know or recognize. This suggests that, in order to elucidate the fundamental principals of the neural computation, we require the system-wide and trans-scale understanding of large neuronal ensembles, connecting molecules, cells and the system.

In the modern biology, massive observation of the complex system is the central driving force. It is exemplified by the long list of "omics" approaches found in the present day: Genomics, proteomics, transcriptomics, connectomics and so on. Those omics approaches have been successful in picturing the landscape of the complex biological systems or phenomena. Starting from the landscape obtained thereby, scientists were often given new ideas to delve into the particular details. Naturally, scientists have employed the omics approaches to understand the neuronal systems, and the brain-wide map (or atlas) have been created for various organisms. These brain-wide map typically describe the gene expression at various locations, as well as the connectivity of the neurons across the regions. Although it is not trivial to predict the dynamical behavior of the brain from these static pictures, having a comprehensive cellular-resolution brain map is the foundation of successful neuroscience research and thus of enormous value.

In this thesis paper, I will describe the new computational framework to collect and analyze brain-wide gene expression and neural connections of the mouse brain. Furthermore, I will present some of the whole-brain analysis results obtained by the proposed method.

In this chapter, I will overview the background theories and experiments, and identify the unsolved challenges. I will begin this chapter by reviewing some of the existing whole-brain mapping data sets and their contributions to the neuroscience (chapter 1.1). Then, I will describe the new experimental techniques called "tissue clearing" and "light-sheet fluorescence microscopy", which, when combined, enable to collect high-quality 3D brain images at much higher throughput than previous approaches (chapter 1.2). These sections are highly important, because present paper describes a novel software to analyze the brain image data obtained by tissue clearing and light-sheet microscopy. Next, I will overview the existing computational methods to analyze the 3D brain images (chapter 1.3). Lastly, I will postulate the software challenges in the tissue clearing-based brain mapping researches, and introduce the problems that will be addressed in the present paper (chapter 1.4).

1.1 Whole-brain mapping of the mouse

In neuroscience research, the mouse (*Mus musculus*) is one of the most intensively studied animal, due to the high genetic and physiological relevance with human and the availability in the laboratory (Laurent, 2020). Although the mouse brain anatomy has been studied over a century, the efforts to generate comprehensive and digital brain atlas was initiated after early 2000s, when megapixel-scale digital image sensors and powerful computer hardware capable of processing massive images were made available. In parallel, advances in genetic manipulation, as well as the viral gene delivery techniques, offered opportunities to label specific cells in the brain.

In this section, I will visit the previous studies attempting to construct the comprehensive brain map of the mouse, which will give the foundation of the present study. I should note that there are primarily two goals of brain mapping: mapping of the gene expression and mapping of the neural connectivity. The former (discussed in chapter 1.1.1) defines the cellular diversity within the brain and links the genotypes and the neural functions. The latter (discussed in chapter 1.1.2) forms the structural foundation underlying neural computing.

1.1.1 Mapping of the gene expression of the whole mouse brain

Fundamentally, it would be reasonable to assume that the logic of neural wiring and functions are written in the genome of the organism. In this regard, mapping of the spatial gene expression within the brain is of high importance, because it provides the links between the genes and the neural functions. For example, if a certain gene is known to be a risk factor of psychiatric diseases, researchers can look up the spatial expression of that gene and identify the candidate brain regions which may be responsible for the phenotypes. This example suggests that gene expression map is a highly useful platform where researchers can mine potentially interesting research hypothesis. From the opposite perspective, researchers can deepen the insight after they obtain some experimental findings. For example, if an interesting electrophysiological response is obtained from a certain brain region, researchers can look up the gene expression map and evaluate what kind of cells are present in the area. Another benefit may be that theorists are able to construct more realistic brain model by taking into account the real distribution of diverse neuronal and glial subtypes.

Based on these scientific motivations, researchers have generated the spatial gene expression map of the mouse brain using different modalities. Among those, here I will describe three modern and popular approaches. It should be noted that, at the present, no single dataset is complete on its own. Each dataset has its own characteristics (such as coverage of the brain areas, the variety of the genes targeted, and spatial resolution), and researchers would need to navigate themselves to an appropriate dataset depending on their purpose.

The first approach is to use *in situ* hybridization (ISH). ISH uses nucleic acid probes having complementary sequence to that of the targeted gene. ISH labeling can target virtually any genes by synthesizing DNA or RNA probes, and thus has a great flexibility. Using this approach, the first comprehensive gene expression map of the mouse brain was reported by Allen Brain Atlas project (Lein et al., 2007). In this project, researchers developed robot-assisted high-throughput pipeline to prepare serial tissue sections and perform ISH staining and imaging. As a result, they obtained the brain-wide spatial expression of over 20,000 genes. Using this massive dataset, they delineated over 1,000 distinct brain regions (Dong, 2008), which has now become the standard in the mouse neuroscience research. Furthermore,

they made the resulting dataset available online, prompting researchers to explore the data using an interactive viewer or programmatically though Allen Brain Atlas APIs (Lau et al., 2008; Ng et al., 2009). Such data scientific environment facilitated the follow-up studies which reported more advanced and detailed analysis (Thompson et al., 2008; Erö et al., 2018). Following the first success of the Allen Brain Atlas, the brain-wide ISH data at several developmental stages was reported (Thompson et al., 2014).

The second approach is by the use of Cre driver transgenic mouse. In this approach, Cre recombinase is inserted in the mouse genome under the promoter of the targeted gene, and the expression of the fluorescent protein (FP) is activated upon Cre-mediated recombination. In contrast to ISH, this approach requires a generation of transgenic mouse for each gene to be investigated. However, since the tissue is endogenously labeled with fluorescent proteins and no post-staining is required, it offers unique advantages in imaging. Using this method, Kim et al revealed the whole-brain map of three major subtypes of gamma-aminobutyric acid (GABA)-ergic neurons, which express parvalbumin (PV), somatostatin (SST) and vasoactive intestinal peptide (VIP), respectively (Kim et al., 2017). Using similar approach, the whole-brain map of cholinergic neurons was recently reported (Li et al., 2018).

The third approach to reveal the spatial gene expression pattern is by the single-cell RNA sequencing (scRNA-seq) method. Here, thousands of individual cells are extracted from the tissue, each attached with the original coordinate in the tissue, and the mRNA expression of each cell is quantified by the next-generation sequencer. This approach offers widest coverage of the genes (essentially all genes in the genome) in one experiment. However, because the throughput of the experiment is determined by the speed of the sequencer, the number of cells that can be quantified (and hence the spatial coverage) is limited. So far, comprehensive scRNA-seq data set of the mouse cortex (Tasic et al., 2016; Tasic et al., 2018) and hypothalamus (Campbell et al., 2017; Chen et al., 2017) have been reported. Recently, brain-wide census of the scRNA-seq was reported (Zeisel et al., 2018), covering about 1 million cells. As a complementary approach, Ortiz et al., 2020 performed mRNA transcriptomics analysis on a regular grid on a brain tissue (a few hundred micrometer pitch). This approach does not offer information on single-cell expression, but provides a brain-wide spatial coverage.

1.1.2 Mapping of the neural connectivity of the whole mouse brain

Reconstructing the circuit diagram of the brain is critical to understand how information is processed or memorized by the neural system. In particular, in the biological brains, multiple computations are executed in parallel in various brain regions, and the outputs are then integrated in other brain regions. This observation requires researchers to study the neural connectivity at the system scale. When discussing the mapping of the neural connectivity, it should be clarified that the approaches can be categorized into three types: macro-, micro-, and meso-scopic approaches.

Macroscopic approaches aims to reveal the global information flow in the brain. Technically, this is usually done by diffusion magnetic resonance imaging (MRI) tractography methods. Although this approach can only capture the bundles of long projecting axons, the technique can scan the entire 3D brain. Using this method, a macroscopic connectome map of the mouse brain was reported (Calabrese et al., 2015).

Microscopic approaches aim to reconstruct the neural connection with nanometer resolution, and the neuronal circuitries are described with synaptic detail. Usually, to acquire such high-resolution images, electron microscopy (EM) is used. Because the throughput of the electron microscope is limited, the scanning of the whole-brain poses a significant challenge. Even though the nanometer-resolution connectome of nematode (*C. elegans*) was achieved as early as in 1986 (White et al., 1986), it was not until the late 2010's when the EM-based connectome of the Drosophila brain was reported (Zheng et al., 2018; Scheffer et al., 2020). Compared to the Drosophila brain (0.01 mm^3 in volume), the adult mouse brain is larger by four orders of magnitude (about 400 mm^3 in volume), and thus, the connectomic analysis using EM have been done only in the partial mouse brain areas (Kasthuri et al., 2015; Motta et al., 2019). To overcome this limitation, a novel methodology was recently proposed, where a use of lattice light-sheet microscopy with physical sample expansion enabled $\sim$ 50 nm resolution (Gao et al., 2019), which may allow to scan large brain tissues with synaptic resolution by using light microscope.

The last approach, the mesoscopic connectomics analysis, is most relevant to the present paper. The first comprehensive mesoscopic connectomics analysis of the mouse brain was reported in 2014 by the Allen Mouse Brain Connectivity Atlas project (Oh et al., 2014; Harris et al., 2019). In this study, they used adeno-associated viral vectors (AAVs) carrying green fluorescent protein (GFP) and injected AAVs to over 300 sites in the brain. The infected neurons express GFP in the soma and axon, which visualizes the projection from the infected area. This approach is called mesoscopic, because the connectivity is quantified by the fluorescence intensity, reflecting the density of the axons. The density of axon correlates with the actual neural connectivity, but strong axon density does not guarantee that there is a strong synaptic connection (i.e. axon may be just passing through that area without having synaptic contacts). In this regard, this approach is termed mesoscopic, because it can offer semi-quantitative neural connectivity at the whole-brain scale. Other AAV-based brain-wide mesoscopic connectivity mapping datasets were independently generated by Zingg et al., 2014 and by Hunnicutt et al., 2014.

Use of glycoprotein gene-deleted rabies viral vectors (RVΔG) is another powerful tool to reveal the brain-wide connectivity (Callaway and Luo, 2015). This approach can be positioned somewhere between microscopic and mesoscopic connectome analysis. Once infected to a neuron (called a starter cell), RV spreads via synaptic connections exclusively in the retrograde direction. By introducing the deletion of the glycoprotein, the virus can be engineered so that the viral infection stops when the virus travels across synapse once (Wickersham et al., 2007). By using this method, researchers can selectively label neurons that have direct synaptic input to the starter cells. Brain-wide connectome analysis using RV have been performed targeting a diverse brain regions (Watabe-Uchida et al., 2012; Ährlund-Richter et al., 2019; Yeo et al., 2019).

Yet another approach is to reconstructing the detailed neuron morphology (axons and dendrites) using high-resolution light microscope data. Due to the lower resolution, this approach cannot identify complete synaptic contacts as is done in EM analysis. Instead, the imaging devices are fast enough to cover the entire mouse brain, offering unique and complementary modality to fill the gap between EM-based and mesoscopic connectome analysis (Gong et al., 2016; Han et al., 2018; Lin et al., 2018; Winnubst et al., 2019).

1.1.3 Serial section-based imaging devices for whole-brain mapping

Before moving on, I would like to dedicate a small section to describe the imaging devices used to acquire the whole mouse brain data sets described in the previous two sections. Because the visible light (400 nm to 700 nm in wavelength) does not penetrate deep into the tissue (only a few hundred micrometers), to acquire 3D image of large tissue such as mouse brain, the tissue need to be physically sectioned.

The most conventional approach is to prepare a consecutive series of thin tissue slices using cryostat machine. This approach is simple and readily reproducible with standard laboratory equipment. However, aligning consecutive tissue slices after microscope imaging is not an easy task due to the mechanical distortions of the tissues. Because of this, the 3D reconstruction artifacts (i.e. discontinuity between slices) are often introduced, which prohibits the accurate registration of the brain with the reference tissue. Another drawback of serial sectioning is that the experimental procedure is laborious and a certain kind of automation must be devised to scale up the data collection. Indeed, a custom-made robot was invented in the initial Allen Brain Atlas project (Lein et al., 2007).

To overcome the limitations of the serial sectioning, two-photon serial tomography (TPST) and the related methods have been invented (Ragan et al., 2012; Zheng et al., 2013; Economo et al., 2016). In this so-called block-face imaging approach, a surface of the 3D tissue is imaged using two-photon or single-photon confocal scanning microscope. Following the imaging, a high-precision vibratome is used to cut a thin slice off the tissue (usually around 50 μm), exposing a new surface for imaging. The whole-tissue scanning proceeds by alternating between imaging and sectioning phases. With this approach, alignment between slices are made much easier allowing more accurate 3D reconstruction. However, due to the fact that the speed of the point-scanning microscopy is limited compared to wide-field imaging, and the finite amount of time is needed to slice the tissue sequentially, this approach suffers relatively long scanning time, often reaching several days for complete scanning of the mouse brain. Due to this limitation, the imaging was often performed with coarse (50-100 µm) z-step size and the missing slices are interpolated computationally (Kim et al., 2017). Recently, FAST method was proposed, where the use of Nipkow spinning disk confocal microscopy significantly improved the imaging speed, at the cost of moderate image resolution (Seiriki et al., 2017). Block-face imaging techniques also lack the accessibility to many post-staining methods, such as ISH and immunostaining. This is because the diffusion of the nucleic acid probes or antibodies into the tissue proceeds very slowly, several orders magnitude slower than the imaging and sectioning speed. Hence, in most of the block-face imaging applications, the specimen needs to be genetically labeled using fluorescent proteins.

1.2 Tissue clearing and light-sheet fluorescence microscopy

As I discussed in the previous section, the imaging methods based on physical sectioning is a very powerful imaging technique to acquire 3D brain images. However, several inherent limitations of these methods were also outlined, which partly explains why such imaging instruments have not spread in the research field. To overcome these limitations, a fundamentally different approach must be taken. In recent years, one of the most promising approach is to use tissue clearing and light-sheet microscopy.

Tissue clearing literally means chemical treatments to transform the biological tissue into a transparent material so that the light can travel through the tissue

without being optically disrupted (i.e. refraction and scattering), while keeping the molecules of interest (e.g. proteins and nucleic acids) intact. The original idea of tissue clearing was pioneered by German anatomist, Dr. Spalteholz, in early 20th century (Spalteholz, 1914). The power of tissue clearing for biological imaging applications became recognized about a hundred years later, by the work by Dodt et al in 2007 (Dodt et al., 2007). The innovation by Dr. Dodt was the idea of combining tissue clearing with light-sheet fluorescence microscope (LSFM), another old technique born in early 20th century[1]. The Renaissance of hundred-years-old techniques embracing the current digital and biological technology offered unparalleled data acquisition throughput of 3D tissues with resolutions that can easily resolve single cells.

Tissue clearing and LSFM have been extensively used in the present paper to collect whole-brain images. Therefore, in the following sections, I will review the technical principles of the tissue clearing methods (chapter 1.2.1) and light-sheet microscopy imaging (chapter 1.2.2), along with some of the milestone applications (chapter 1.2.3).

1.2.1 Tissue clearing techniques

After the work by Dr. Dodt, a significant amount of researches have been carried out to improve the performance of the clearing methods, and dozens of protocols have been reported thus far. As more and more chemicals effective in tissue clearing have been identified, the researchers are gaining a general understanding of the tissue clearing chemistry (Tainaka et al., 2016; Ueda et al., 2020).

The light gets scattered or refracted in the tissue because of the inhomogeneous refractive index (RI) in the intra- and extracellular compartments. Namely, the tissue is mainly consisted of lipid bilayer (RI $\simeq$ 1.45), water (RI = 1.33) and protein molecules (RI $\simeq$ 1.50 to 1.55). Light propagation within the tissue is disrupted at the boundaries of these cellular structures. Therefore, the key concept of tissue clearing is homogenizing the refractive index in the tissue.

In so-called hydrophobic clearing methods (such as BABB (Dodt et al., 2007), 3DISCO (Ertürk et al., 2012) and iDISCO (Renier et al., 2014)), the water in the tissue is removed (dehydration) by immersing the tissue in organic solvents. Subsequently, the tissue is immersed in another organic solvents to elute the lipids (delipidation), rendering the tissue essentially a cross-linked protein gel. Lastly, the tissue is immersed in high-RI medium (RI $\simeq$ 1.55) so that the refractive index of the tissue environment is matched with that of the protein (RI matching). These chemical treatment is able to make relatively large organs (e.g. the mouse brain) almost completely transparent. However, especially in the early days[2], many of the hydrophobic clearing methods quenched the signal of the fluorescent proteins. In addition, some of the organic compounds were hazardous and needs careful handling in the laboratory.

To address these problems, researchers later developed so-called hydrophilic clearing methods (such as Sca*l*e (Hama et al., 2011), SeeDB (Ke, Fujimoto, and Imai, 2013) and CUBIC (Susaki et al., 2014)). In hydrophilic clearing methods, the lipid

[1] The original form of the light-sheet microscope (then called "ultramicroscope") was invented by Zsigmondy and Siedentopf in 1902. Zsigmondy won the Nobel prize in chemistry in 1925 for the invention of ultramicroscope. Before introduction to biological imaging, ultramicroscope has been used in physics experiments to, for example, observe the Brownian motion of the colloidal particles suspended in 3D space.

[2] Recently, researchers have developed organic solvent based clearing methods that are able to preserve or recover the fluorescent proteins signals, such as FDISCO (Qi et al., 2019) and vDISCO (Pan et al., 2019).

in the tissue is removed by the water-soluble chemicals, such as detergents and aminoalchohols. Urea may be used to enhance the permeability. After delipidation is complete, the tissue is immersed in RI matching solution. RI matching solution typically contains molecules with high water solubility and high polarizability (such as sugar and contrast agents), which typically reaches RI value between 1.45 to 1.52. In general, hydrophilic clearing is able to minimize the quenching of the FPs, allowing genetical or viral labeling of cells.

Soon after the invention of the hydrophilic reagents, another key element of tissue clearing was realized: the decolorization (Tainaka et al., 2014). In the mouse brain and body, heme is the most dominant light-absorbing compound. With the presence of heme, light cannot travel deep into the tissue even if the effective RI homogenization is performed. At present, several chemical compounds are known (such as aminoalcohols) that can effectively elute the heme and other pigmentation (Tainaka et al., 2014; Zhao et al., 2020; Pende et al., 2020).

It is also important to mention the importance of tissue fixation in the context of tissue clearing chemistry (Gradinaru et al., 2018). By applying appropriate fixation method, one can effectively keep the molecules of interest intact, while removing other unwanted molecules. The most popular fixation method is the application of paraformaldehyde (PFA). PFA efficiently reacts with the amino acids and cross-link the proteins in the tissue, while lipid components do not react. Therefore, one can wash out the lipid components while keeping the proteins during delipidation step. Although PFA fixation is highly suitable for many of the clearing methods, one can engineer the fixation chemistry to further enhance the clearing performance. Such approach is called hydrogel-based clearing methods, and the idea was pioneered by CLARITY method (Chung et al., 2013). In CLARITY method, the tissue was embedded in acrylamide polymer network. This enabled the tight fixation of the small chemicals (such as mRNA and nuerotransmitters) that would otherwise be washed out. Further, hydrogel embedding enhanced the structural rigidity of the tissue, allowing to use rather harsh conditions to accelerate the delipidation using heat or electric fields (Chung et al., 2013; Tomer et al., 2014). Later generations of tissue-hydrogel engineering (such as MAP (Ku et al., 2016), SHIELD (Park et al., 2019) and ELAST (Ku et al., 2020)) further enhanced the flexible manipulation of the tissue's mechanical, chemical and physical properties.

Lastly, I will discuss the labeling methods for the cleared tissue. Many of the latest tissue clearing methods nicely preserve the signals from fluorescent proteins (FPs). Thus, if a reporter mouse line is available, researchers can reliably image the labelled cells. In addition, expression of the FP may be introduced by the injection of AAV or RV. Other powerful labeling approach is the immunostaining (Renier et al., 2014; Susaki et al., 2020; Zhao et al., 2020). Intriguingly, the tissue homogenization steps in tissue clearing increases the permeability of the tissue, so that relatively large macromolecules such as antibodies can diffuse into the tissue more rapidly than the untreated one. In the hydrogel-based methods, the polymer-enhanced tissue may also be beneficial in transporting the fluorescent probes or antibodies into the deep tissue using pH, heat or electric fields (Kim et al., 2015a; Murray et al., 2015; Yun et al., 2019). The last frontier of the labeling methods for tissue clearing would be ISH labeling. Although ISH labeling of cleared tissue has been demonstrated in a small block of tissues (Chung et al., 2013; Park et al., 2019), the generally applicable method for whole-brain ISH labeling is yet to be demonstrated.

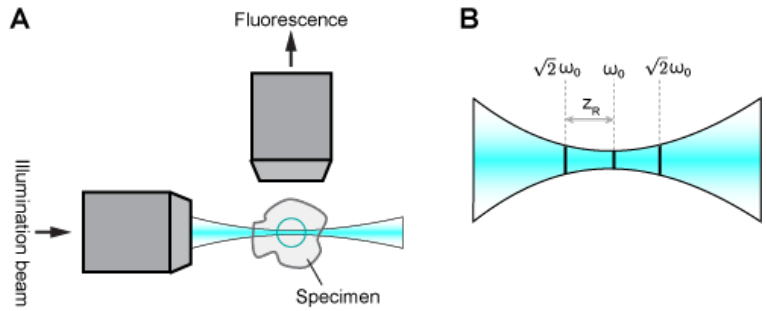

FIGURE 1.1: Schematic of the light-sheet fluorescence microscope (LSFM)

A. The illustration showing the standard configuration of the LSFM. **B**. The illustration of the beam wait diameter (ω_0) and the Rayleigh range (Z_R).

1.2.2 Light-sheet fluorescence microscopy

Once the tissue is cleared, one can use light microscope to observe the 3D tissue. Conventionally, to acquire 3D fluorescence images, the confocal microscope (Pawley, 2006) has been, and still is, used. For the purpose of large tissue imaging such as mouse brain, however, the light-sheet fluorescence microscopy (LSFM) (Power and Huisken, 2017) was proved to be an extremely powerful solution.

In the LSFM, the specimen is illuminated by a thin sheet of laser light (typically 1 μm to 10 μm thickness), and the fluorophores located in the illuminated plane is selectively excited (Figure 1.1 **A**). The emitted fluorescence is then collected by the second objective lens, which is orthogonally positioned with respect to the illumination lens, and imaged onto a digital image sensor (Figure 1.1 **A**). In contrast to the point scanning approaches used in the confocal microscopes, the LSFM collects a two-dimensional image in one capture, instead of a point. This allows massively improved image acquisition speed. Indeed, with the latest techniques, the whole mouse brain scan can be completed within 10 minutes (in the case of macroscopic scan ($\sim$ 6 µm)) (Voigt et al., 2019) to several hours (in the case of high resolution scan ($\sim$ 0.6 µm)) (Tomer et al., 2014; Matsumoto et al., 2019).

There are essentially two types of implementations of LSFM. In the first approach, a laser sheet is generated by focusing a beam with a cylindrical lens. This approach is often called the ultramicroscope (terminology used by Zsigmondy) or selective plane illumination microscope (SPIM) (Huisken et al., 2004). The second approach is called digital scanned laser light-sheet fluorescence microscopy (DSLM) (Keller et al., 2008; Silvestri et al., 2012; Tomer et al., 2014). In DSLM, one-dimensional laser beam is rapidly scanned vertically using a galvo mirror to generate a virtual plane of light. Both approaches have their own advantages and disadvantages (for more information, see the review by Power and Huisken, 2017).

In LSFM, the laser sheet is not a perfect plane, but rather is a focusing beam. Because a focusing beam has an axially stretched profile, one can effectively regard the region near the beam waist as a homogeneous sheet (Figure 1.1 **A**, **B**). Given the beam waist diameter ω_0 (defined by the distance where the intensity becomes $1/e^2$ of the peak), the effective width of the sheet is characterized by the Raylegih range, Z_R, where the beam diameter is equal to $\sqrt{2}\omega_0$ (Figure 1.1 **B**). If the beam is a normal Gaussian beam, the relationship between beam waist ω_0 and the Rayleigh range Z_R

is given by the following formula (Power and Huisken, 2017):

$$Z_{\mathrm{R}} = \frac{\pi\omega_0^2}{\lambda} \tag{1.1}$$

where λ is the wavelength of the light. This means that, as the light sheet is made thinner, the effective area usable for light sheet decreases quadratically. This trade-off between the axial resolution and the field of view (FOV) is a very important consideration in designing LSFM.

To overcome this trade-off, several optical techniques have been proposed. One of the idea is to use non-Gaussian beam which has more elongated profile along the propagation direction, such as Bessel (Gao et al., 2014), Airy (Vettenburg et al., 2014) and lattice (Chen et al., 2014; Chang et al., 2019) light sheet. The other, and very simple, approach is tiling several light-sheet whose focus is shifted by the Rayleigh range (Gao, 2015). A "continuous" version of the tiled light sheet approach is called the axially swept light sheet microscope (ASLM) (Dean et al., 2015; Chakraborty et al., 2019). In this setup, a focus position of the light sheet is continuously swept along the propagation axis. The sweep of the light sheet focus is synchronized with the rolling shutter of the sCMOS camera, which allows to reject the out-of-focus signal. In the present paper, some data were acquired by tiled-light sheet microscope, and others were acquired with ASLM-based microscope (see chapter 2.1.3).

1.2.3 Applications of tissue clearing to whole-brain/body imaging

In the previous two chapters, I have reviewed the tissue clearing and LSFM imaging methods. By combining these novel technologies, a handful of innovative applications have been demonstrated thus far.

In terms of the circuit mapping, researchers used CLARITY method together with rabies virus injection to identify the brain-wide input to the dopamine neurons in the ventral tegmental area (VTA) and substantia nigra, compact part (SNc) (Menegas et al., 2015). In another study, a long-range projection from paravalubmin-positive neurons in the globus pallidus, external segment (GPe) have been reconstructed at single-cell level using SHIELD method (Park et al., 2019).

One of the broadly useful application of tissue clearing is the imaging of immediate early genes (IEGs), such as c-Fos, Arc and Egr1. The expression of the IEGs are enriched in the neuron with high activity having frequent firing (Sagar, Sharp, and Curran, 1988). Thus, IEG expression amount can be used as a proxy to reconstruct the neural activity profile from the post-fixed brain tissues. The expression amount of IEGs can be visualized at whole-brain scale by genetically introducing fluorescent proteins or by post-staining with antibodies. Brain-wide quantification of the IEGs have been performed under various experimental conditions, such as light exposure (Susaki et al., 2014), drug administration (Tatsuki et al., 2016; Renier et al., 2016; Salinas et al., 2018) and behavioral stimulation (Renier et al., 2016; Roy et al., 2019). These researches have successfully identified neural clusters responsible for a particular brain function.

Another important branch of tissue clearing is imaging of the peripheral nervous system. Because the sensory and motor axons extends over a long distance, a 3D volumetric observation and quantification via tissue clearing is crucial. Some clearing methods are capable of transforming the bones clear, allowing the whole-spine or whole-body imaging of the nervous system (Tainaka et al., 2014; Yang et al., 2014; Renier et al., 2014; Greenbaum et al., 2017; Jing et al., 2018; Cai et al., 2019).

Brain vasculature is of significant interest in neuroscience, since neurons are one of the most energy-demanding cells in the body and an efficient and dynamically tubable energy delivery is essential. Because the blood vessels form a complex 3D structure, the study of vascular system was enormously challenging with conventional 2D slice methods. Motivated by this challenge, researchers have recently developed methods to image 3D vascular structure by tissue clearing and chemical labeling of the veins and arteries (Kirst et al., 2020; Todorov et al., 2020; Miyawaki et al., 2020). Using these methods, a complete atlas of mouse brain vasculature have been created (Kirst et al., 2020; Todorov et al., 2020).

Although applications of tissue clearing is most developed in mouse brain researches, the methods can be applied to other organs or other species with minimal modification of the protocols. So far, researchers applied tissue clearing to non-human primate brains (Susaki et al., 2014), human brains and other organs (Inoue et al., 2019; Zhao et al., 2020), along with other vertebrate or invertebrate organisms (Pende et al., 2020).

As an important milestone in tissue clearing technique advancements, the researchers from Ueda group reported the construction of CUBIC-Atlas (Murakami et al., 2018). Here, the researchers combined high-performance and expansion-assisted tissue clearing, high-resolution LSFM imaging and high-throughput image analysis to digitally record all of the cell nuclei present in the mouse brain (amounting to approximately 100 million cells). As a result, a single-cell resolution map of the 3D mouse brain named CUBIC-Atlas was generated. CUBIC-Atlas serves an central role in the present paper (chapter 3.1.1).

1.3 Automatic analysis methods for 3D brain images

Using the tissue clearing and LSFM imaging described in chapter 1.2, researchers can rapidly acquire 3D brain images with single-cell-resolution. To reach biological discovery, however, those images must be extensively and carefully analyzed. Importantly, the dataset from whole-brain scanning is massive; in the case of mouse brain imaging, the raw data amount in uncompressed 16 bit image format ranges from 15 GB (in the case of macroscopic scan ($\sim$ 6 µm)) to 10 TB (in the case of high resolution scan ($\sim$ 0.6 µm)). If multi-color imaging is performed, these numbers are multiplied by the number of the chanels. Hence, manually annotating and analyzing the entire image is practically not possible, and an automated image analysis routine must be developed to gain quantitative understanding. In this section, I review the existing algorithms and software used in 3D brain mapping applications. Reflecting the common analysis steps of the cleared tissue imaging, the section divides into three parts. First, I review the cell detection/segmentation methods used in the cleared tissue imaging (chapter 1.3.1). Then I visit the brain registration algorithms to accurately align multiple brains (chapter 1.3.2). Finally, I will mention the use of cloud computing platforms to accelerate the analysis of big image data (chapter 1.3.3).

1.3.1 Cell detection and segmentation

The first step in image analysis is identifying the objects of interest (e.g. cell body, cell nuclei, axons or dendrites). In some applications, knowing the total number of the objects may be sufficient (object detection), while in other applications, a detailed morphology of the object may be necessary (object segmentation). Although such

object detection/segmentation task is very basic in biology, when it is applied to whole tissue scale images, it presents significant challenges.

The first difficulty lies in the fact that the object morphology and brightness as well as the background noise levels varies depending on the brain region. To overcome this, one often needs to design a customized set of image filters and thresholding scheme optimized to a specific data set (Renier et al., 2016; Murakami et al., 2018; Kirst et al., 2020). To avoid the need to manually design a complicated image processing routine, the unsupervised learning approach can be used (Amat et al., 2015; Matsumoto et al., 2019). If a sufficiently large training data set is available, supervised learning approaches, such as deep convolutional neural networks (CNN), may be used (Kim et al., 2015b; Pan et al., 2019).

The second challenge is the massiveness of the image data. Modern microscopes for whole tissue scanning can easily produce terabytes of images within a few hours (Wan, McDole, and Keller, 2019). To keep up with such enormous data production speed, image analysis program must be optimized and parallelized. Hence, many programs now embrace the power of graphic processing units (GPUs) (Amat et al., 2015; Murakami et al., 2018).

1.3.2 Brain registration

The second important step is brain registration. Here, registration means a computation of transformation to align the shape of one brain (moving image) with that of the reference (fixed image). Because tissues or organs can elastically deform their shape, linear (i.e. rigid or affine) transformation is not enough to correct for the individual differences, and thus a non-rigid transformation must be considered. By running registration, brains from different experiments/subjects can be virtually overlaid together, allowing voxel-by-voxel comparison and quantification. The algorithms for 3D image registration is most advanced in medical or clinical applications, which deals with 3D images of human organs such as brains acquired by magnetic resonance imaging (MRI) or computed tomography (CT) scans. Consequently, these algorithms are adopted and optimized for the brain registration of other organisms, such as mouse.

Mathematically, the 3D deformable registration is formulated as follows. Let F and M denote the fixed and moving image, respectively, and ϕ represent the mapping of the voxel coordinates from M to I. The deformable image registration attempts to find ϕ^* which minimizes the cost function:

$$\phi^* = \arg\min_{\phi} \{-\mathcal{L}_{\text{sim}}(F, M(\phi)) + \mathcal{L}_{\text{reg}}(\phi)\} \tag{1.2}$$

Here, $M(\phi)$ is image M warped by ϕ, $\mathcal{L}_{\text{sim}}(\cdot,\cdot)$ measures the similarity between two images, and $\mathcal{L}_{\text{reg}}(\phi)$ represents the regularization term to force the ϕ to be smooth.

There are several common formulations for ϕ, $\mathcal{L}_{\text{sim}}$ and $\mathcal{L}_{\text{reg}}$. The popular choice of image similarity metrics ($\mathcal{L}_{\text{sim}}$) includes squared sum of intensity differences (SSD), normalized cross-correlation (NCC), and normalized mutual information (NMI). To enforce smoothness in ϕ, $\mathcal{L}_{\text{reg}}$ is often given by the "bending energy", which is given by the L2-norm of the second-derivatives of the vector field ϕ (Rueckert et al., 1999). Other approach is to compute the Jacobian matrix of the vector field and evaluate the sign of determinant value (Mok and Chung, 2020).

ϕ is often formulated as a displacement vector field. Algorithms based on this formulation includes Demons (Thirion, 1996) and free-form deformation with b-splines (Rueckert et al., 1999). Other approach formulate ϕ as the diffeomorphic

transformation, which is advantageous in ensuring the topology preservation and the exactness of the inverse transformation. Popular diffeomorphic registration methods include diffeomorphic Demons (Vercauteren et al., 2009) and symmetric image normalization method (SyN) (Avants et al., 2008). Furthermore, researchers recently started to explore the possibility of using deep neural networks as a universal function to generate a transformation between images (Balakrishnan et al., 2018; Mok and Chung, 2020). Combined with optimization methods developed in deep neural networks and GPU-accelerated computing, these learning-based approaches offer comparable accuracy with the conventional iterative minimization methods, while significantly reducing the time required to compute the warp.

In the mouse brain mapping literature, two registration algorithms are most popularly used. The ClearMap framework (Renier et al., 2016) uses elastix (Klein et al., 2010), which essentially implements the deformable registration using 3D b-spline. In the CUBIC pipeline (Susaki et al., 2014; Susaki et al., 2015), symmetric diffeomorphic transformation (SyN) algorithm implemented in ANTs library is used (Avants et al., 2008). SyN was shown to be one of the most accurate method in the comprehensive benchmarking studies (Klein et al., 2009; Nazib, Fookes, and Perrin, 2018). However, SyN is computationally demanding and requires long computation time, while B-Spline deformation in elastix is relatively fast in computation and gives moderate accuracy.

1.3.3 Big image data analysis using modern cloud computing

As I have reviewed in the previous sections, the current 3D brain mapping techniques can easily produce multi-terabyte scale image data in single experiment. To keep up with this enormous data production rate, computational analysis, as well as the storage, needs to be scaled accordingly. To pursue scalable and universally available solution, researchers have started to utilize modern cloud computing platforms.

Automated segmentation of cell morphology is of highest demand in many applications. Particularly, advances in deep neural network have provided ever more accurate algorithms. However, applying deep learning models to large image data requires multi-GPU compute environment, and the number and the type of GPU required may be different between training and inference phase. To dynamically and flexibly utilize those hardware resources, cloud computing is suitable than local workstations, and several software solutions have been proposed (Haberl et al., 2018; Bannon et al., 2018; Falk et al., 2019; Wu et al., 2019). Some of these software are designed to be deployed on the public cloud platforms provided by companies like Google or Amazon, rather than the private cloud owned by the research institute. This design choice allows the developer to (1) use platform-specific functionalities and APIs to enrich the features and accelerate the development and (2) allocate virtually unlimited number of CPUs and GPUs to scale the computation. On the other hand, this design choice presents the unique and interesting challenge, which is that the architecture of the program directly determines the cost billed by the public cloud provider. Due to this fact, optimizations are not limited to the algorithm and the code, but also include how the cloud resources should be managed (Bannon et al., 2018; Wu et al., 2019).

Another important use of cloud computing is the storage of the large image data. In addition to the advantage that the storage space in cloud can effectively grow unlimitedly, the more important point is sharing of the data. In the present day, the

network speed is often the bottleneck when transferring large data between laboratories. Rather than copying local data to another location, a more sensible approach is to store data in the cloud, and let many users access the common data (i.e. the model where user *come* to the data, instead of data *go* to the user). To facilitate such scheme, many cloud-native storage system for large array data have been proposed (Kleissas et al., 2019; Katz and Plaza, 2019). In these storage system, the large 3D array is stored in a chunked format, so that a subvolume can be quickly retrieved.

Based on the cloud-native storage system, web-based image viewers have been developed (Saalfeld et al., 2009; Boergens et al., 2017; Dorkenwald et al., 2020). These software allows the user to see the massive image data through a web browser, as well as annotate the images. Image annotation is important to manually segment the object of interest, prepare training data set for machine-learning based image analysis, or proof-reading the automatically generated predictions.

1.4 Towards community-driven mouse brain mapping

As I have reviewed in the previous chapters, tissue clearing and LSFM imaging now offer a novel method to rapidly scan the high-quality 3D brain images with single cell resolution. With these technological advancements, the field is entering an era where whole-brain mapping projects, which conventionally required institution-scale resources and efforts, can be carried out by individual laboratories, or even by a single researcher.

In this regard, I argue that the current technological stage can be thought analogous to the dawn of genome sequencing technology in early 2000s. The initial sequencing of the human genome was achieved by approximately 12 years of collective efforts and over a hundred million dollars of financial investment (Lander et al., 2001; Venter et al., 2001). Now, with the advent of the next-generation sequencer, human genome can be sequenced within a day at around a thousand dollars. Consequently, genome sequencing became accessible to any researchers, and the collection of genome sequences of human as well as other organisms are carried out worldwide, offering opportunities for big data-driven discoveries as we appreciate today.

This presumed parallelism between genomics and brain mapping is the key starting point of the present thesis (Figure 1.2). If this parallelism is assumed, it suggests that brain mapping could potentially follow the same trajectory as what genome science have achieved. Namely, it may be possible to establish a distributed data collection scheme where individual researchers across the globe perform whole mouse brain mapping experiments and share the resulting dataset. Such community-supported brain mapping scheme will accelerate the collection of brain mapping data targeting diverse genes and neural circuits under various experimental conditions, potentially providing substantially wider coverage than the existing, centrally-generated whole-brain mapping datasets such as Allen Brain Atlas.

However, because tissue clearing-based brain mapping is still an emerging technique, hardly any studies have addressed the problem of integration and sharing of the data across laboratories. Consequently, the brain mapping data is usually stored in isolation within the laboratory, and never re-analyzed in later studies. Thus, the main purpose of this paper is to propose a software solution to allow effective integration of whole mouse brain datasets distributedly collected by researchers.

In composing this study, I first clarify the motivations and potential benefits of sharing brain mapping data (chapter 1.4.1). Then I mention previous attempts to

integrate mouse brain anatomy dataset across published studies, and explain why the present approach using tissue clearing-based 3D imaging can offer qualitatively different paradigm (chapter 1.4.2). Next, I discuss the insights learned from other databasing attempts in related discipline of science, including genmonics (chapter 1.4.3) and human neuroimaging using magnetic resonance imaging (MRI) (chapter 1.4.4). Based on these observations, in chapter 1.4.5, I will formulate the challenges that needs to be addressed in designing a software for distributed brain mapping.

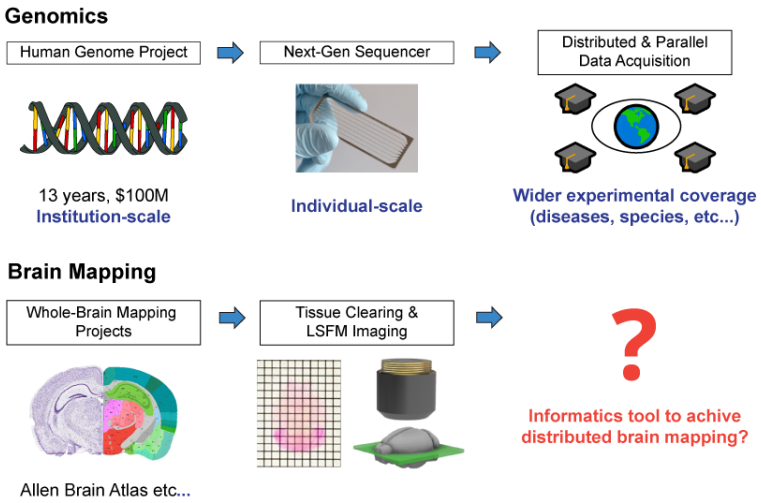

FIGURE 1.2: The proposed parallelism between genomics and brain mapping.

1.4.1 Why share brain mapping data?

To motivate the present study, it is important to clarify why data sharing is important in the whole mouse brain mapping studies.

The first consideration point is the richness and reusability of the whole brain mapping data. Whole brain data is enormously complex, composed of over thousands of 2D images and contains 10^4 to 10^8 targeted cells. Although automatic image analysis methods exist (as discussed in chapter 1.3), it is still practically not feasible to analyze the whole-brain data from all possible aspects. When publishing a paper, the authors would only describe the novel findings relevant in that study, and other potentially interesting observation would remain undocumented. In this sense, I argue that whole-brain dataset is similar to genome sequence data in nature. By depositing the complete brain mapping result in a data sharing platform, it offers opportunities for other researchers to re-analyze the result to address different biological questions.

The second benefit of data sharing is that it allows to collect a massive amount of datasets that are far beyond the scope of a single study, and run a meta-analysis to discover new insights. A successful example of such population meta-analysis can be found in the human neuroimaging field (see chapter 1.4.4), where researchers were able to reconstruct the default mode network (DMN) of the cortex from the population data of resting state fMRI measurement (Biswal et al., 2010). Another example includes the construction of the forward model to predict the brain activation from the stimuli, and the reverse model to predict the stimuli from the brain activation map (Yarkoni et al., 2011; Yannick, Bertrand, and Gael, 2014), using the large

fMRI dataset. Such large-scale meta-analysis scheme would be equally applicable to mouse brain mappings. For example, by aggregating the whole-brain IEG measurements (chapter 1.2.3), one may be able to relate the brain-wide activity profile and the mouse behaviour.

The third benefit is that the data sharing platform allows to compare the gene expression or neural connection under diverse experimental conditions. It is important to note that the existing mouse brain databases such as Allen Brain Atlas usually investigate the brain of the healthy adult wild-type mouse. In many biological researches, however, the brains in perturbed state (such as certain disease, environmental or developmental conditions) offer valuable insights, and thus of the major interest of many researchers. Indeed, in the genomics, the genome data of individuals having drastic phenotypes (such as diseases) have been instrumental in discovering the gene mutations. Because such interests are question specific, centrally-led database would not be able to determine which condition would be of high impact in the research community. Rather, such focused data collection would be driven by a group of researchers having common interest or question. A data sharing platform would encourage such collaborative data collection. A community-driven mouse brain sharing platform, by its construction, would be able to offer the dataset that are highly demanded by the researchers.

1.4.2 Previous attempts to integrate mouse brain anatomy data

As I have reviewed in chapter 1.1, most of the currently existing mouse brain database were generated by large projects, such as Allen Brain Atlas (Lein et al., 2007; Oh et al., 2014), Brain Architecture project (Kim et al., 2017) and MouseLight project (Winnubst et al., 2019). These databases are led by a single or multiple leading institutions, and did not offer the opportunities for external researchers to submit data. The efforts to integrate the anatomy data from different studies are surprisingly unexplored in this field, largely due to the lack of motivation to do so, and the difficulty of mapping slice image data to the 3D reference brain space. Nonetheless, such concept was recently explored independently by Fürth et al., 2018 and by Chen et al., 2019. These frameworks both assume that a series of tissue slice images are given as input, and the images are aligned with the reference brain to allow accurate integration of multiple brains. These frameworks also offer simple web services to share the resulting data. However, the web service is still at proof-of-concept stage (e.g. minimal or no graphical user interface and undocumented set of APIs) and needs further developments to be used as a foundational data repository that serves the community.

In the present paper, I assume that the input is given as a 3D whole mouse brain image, a novel scheme which has never been explored before. This design choice offers unique advantages qualititatively different from the previous tissue slice-based approaches. First, by having the complete 3D image of the brain, the mapping of the brain is made more accurate and easier, because the registration can take into account the 3D structural context. Second, by restricting to only whole brain data, the database records are made more homogeneous, offering better experiences for the future data mining projects. Thirdly, having the complete 3D image of the brain means that the dataset contains the "internal control". To illustrate this, let us consider the brain activity reconstruction using IEGs (as discussed in chapter 1.2.3). If a researcher combines multiple datasets from several brain mapping studies, and attempts to identify the brain region affected by a certain stimuli, analysis must also

consider the regions that should not be affected (control regions). Having the complete 3D image means that one can reliably test those null hypothesis, and ensure the choice and quality of the data. Further, one can normalize the signal strength by looking at the distribution of the whole data. Indeed, such analysis is very common in the meta-analysis of human fMRI studies (see chapter 1.4.4). If a partial brain data is given, such verification and normalization is not possible, and data miner cannot eliminate the possibility of the experimental artifacts.

1.4.3 Database framework for genomics

To design a data sharing framework for whole-brain mapping, it is fruitful to review the database frameworks developed in genomics, where the community sharing of the data is among the most advanced and successful.

The core infrastructure for the collection and archiving of the sequence data is organised by International Nucleotide Sequence Database Collaboration (INSDC) (Arita, Karsch-Mizrachi, and Cochrane, 2020), which is a international collaboration of the DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (ENA) and GenBank. As of 2020, INSDC hosts over 9 petabytes of data contributed by the research community. All sequence data in the database (except for the studies involving human privacy) are publicly available through the internet access without use restrictions or licensing requirements. Individual sequence data submitted to the database are given a unique and permanent accession number, which can be used to reference the data in the published literature. In addition, all sequence data is accompanied with the reference to the study and the identity of the data depositor to keep track of the source of information.

Building upon this database, a variety of web-based tools are provided to query, analyze and visualize the data. For instance, BLAST (Altschul et al., 1990; Altschul et al., 1997) provides sequence similarity search, which rapidly scans through massive database and finds the similar sequence. Another frequently used tool is GenomeBrowser (Karolchik, Hinrichs, and Kent, 2009), which provides an interactive visualization tool to display the genome sequence accompanied by a series of annotations.

Since its launch, genome databases have served central role in biological and medical researches with a wide spectrum of use cases. In the light, day-to-day use cases, researcher would search whether the certain domain in the protein amino acid sequence is conserved across species, which often indicates the functional center of the protein. In the most data-intensive use cases, researchers would run genome-wide association studies (GWAS) using the massive dataset of human genome, to discover the mutations that is linked to a certain phenotype, such as diseases.

The insights that are applicable to mouse brain database are several fold. First, the success of genome database may be attributed to the design where sequence alignment and search are offered as a cloud computing service. By having this design, all deposited sequence data are processed and indexed using the same algorithm, which ensures the consistency in the database. Further, by allowing users to run complex query in the cloud, users do not need to download the large data to the local computer. Second, the development of web-based tools that allow to quickly analyze and visualize the data was a quite successful effort. Web-based tools alleviate the need to install specialized software on the user's computer. Further, web-based tool allows to share the analysis results with collaborators by sharing the link (Karolchik, Hinrichs, and Kent, 2009). These features greatly enhance the user experience, especially for the biology experts who may not have extensive knowledge on informatics.

1.4.4 Database framework for MRI-based human neuroimaging

Genomic database platforms provide one of the ideal forms of the scientific data sharing. As a more directly related discipline to the present study, next I review the data sharing efforts in the human neuroimaging field.

Since the very early days of the human neuroimaging using MRI instruments, the importance of the data sharing was recognized in the field. Consequently, several database platforms to store the raw image data have been developed, including fMRIDC (Van Horn et al., 2001), INDI (Mennes et al., 2013), OpenfMRI (Poldrack et al., 2013) and connectomeDB (Hodge et al., 2016). Common in all platforms, individual submitted brains are aligned with the reference brain, and are given a unique ID so that the data can be referenced from the published literature. In addition to the storage of the raw image data, there are several platforms that aim to collect "curated" neuroimaging dataset, such as SumsDB (Dickson, Drury, and Van Essen, 2001), NeuroValut (Gorgolewski et al., 2016) and BALSA (Van Essen et al., 2017). The "curated" data here means some kind of processed MRI data and includes a diverse data formats. For example, NeuroVault is designed to collect unthresholded voxel-wise statistical maps showing the activated or repressed areas extracted from the fMRI measurements. BALSA, on the other hand, is designed to accumulate any kind of analysis results generated by Connectome Workbench software.

The raw image database and curated database are complementary in terms of its mission and usage. Fundamentally, the raw image data is the source of all research results, so the sharing of it is essential. However, fMRI studies often produce hundreds of gigabytes of image data in one study, placing a large load on the database management as well as the database user. Furthermore, re-analysis of the raw image data (especially task-based fMRI data) can be quite laborious, often requiring the precise knowledge of the experimental design. Curated data, on the other hand, allows the data miner to skip the tedious handling of the raw images, and are often more favorable in situations involving meta-analysis of the MRI data.

Compared to genomics, the data sharing of the human neuroimaging community is rather dispersed, and the efforts in the data integration are still ongoing. This is partly because the human neuroimaging dataset are more diverse in the format than genome sequences. Especially in cognitive studies using fMRI, the diversity of the format comes from the complicated design of the tasks, where the task instruction, stimuli and repose are different. To overcome this diversity problem and achieve a coherent data organization, the importance in the common ontology to describe the experimental design is being recognized, and the ontology sets specialized in human cognitive studies have been proposed, such as CogPO (Turner and Laird, 2012).

Despite the challenges in exchanging diverse neuroimaging data, the neuroimaging databases have been used in many studies to derive novel discoveries. For example, by integrating hundreds of human brain imaging data, the spontaneous brain activity at the resting state (also called the default mode network) was decoded (Biswal et al., 2010). Integration of many brain data is also indispensable in generating ever more precise parcellation of the human brain (Glasser et al., 2016). In other studies, researchers have integrated fRMI datasets from serveral studies and constructed a model to predict the neural response from the stimuli (forward inference) or predict the stimuli from the neural response (reverse inference) (Yarkoni et al., 2011; Yannick, Bertrand, and Gael, 2014).

There are several insights that may be instructive in designing database for mouse brain mapping. First, a separation of raw image database and the curated database

would be reasonable. Second, it is easy to foresee that the mouse brain mapping would also face the challenges in the diversity of the dataset. As discussed in chapter 1.1, there are diverse experimental methods to label the gene expression or neural connectivity. Thus, an organized framework to describe the sample information as well as the experimental conditions would be essential.

1.4.5 Software challenges towards community-driven mouse brain mapping

Based on the observations in the previous section, here I formulate the software challenges and requirements that needs to be addressed in designing a data sharing platform for whole brain mapping.

1. Standardization of the data

   In order to create a homogeneous collection of brain mapping data, the data standardization procedure must be meticulously defined. This involves the choice of the reference brain, and the mapping strategies to register individual brain to the reference. Further, the input data format must be rigidly defined. Definition of the input data format is deeply linked with the scope of the data that the platform aims to collect.

2. Web-based interface

   This is perhaps quite obvious in the present day, but the platform should be constructed in the cloud space to provide superior accessibility through the internet. Important consideration is that the system should offer both programatic and graphical interface to serve different type of users (as discussed in chapter 1.4.3 and 1.4.4). For the light use cases, the framework should present graphical user interfaces (GUIs) to provide interactive user experience. On the other hand, for the users attempting deep data mining, or for the integration with third party applications, the framework should offer programmatic access to the service.

3. Cloud-based end-to-end analysis

   As discussed in chapter 1.3, the tissue clearing-based whole brain mapping can produce over hundreds of gigabytes image data. Due to this massive size, the analysis poses significant challenges to most of the experimentalist, who do not necessarily have expertise in high-performance computing or powerful computer resources. For this reason, I postulate that the software should provide the entire data analysis workflow, not just the service for data sharing. This approach is modeled after the ecosystem offered by BALSA and Connectome Workbench software in human brain imaging. Connectome Workbench provides an integrated analysis environment for the human MRI data, and it seamlessly connects with BASLA platform to share the analysis results.

4. Scalability of the system

   In designing the system, the scalability must be carefully considered. Although the number of users may not be large at the initial stage, the number could grow rapidly in the future. Indeed, according to the PubMed search, between 2019 to 2020, 193 papers have been published which contain the word "tissue clearing" in the title or in the abstract[3]. These papers can be the potential contributers to the data sharing platform. Thus, the system should be able to

flexibly grow to support hundreds of users, without needing to rewrite the server program.

5. A comprehensive description of the data

   As discussed in the human neuroimaging databases, labeling each data entry with organized and consistent description is the key to the successful data integration. The necessary tags, and potentially the appropriate ontology, must be carefully designed.

In the following chapters, I will present the software implementation that addresses the requirements outlined above, named CUBIC-Cloud (chapter 3). Some aspect of the system is still experimental, but it offers a set of functionalities to provide a novel cloud environment to analyze and share whole mouse brain data.

Further, I will demonstrate the usability of the proposed CUBIC-Cloud framework in a variety of neuroscientific applications (Chapter 4). The demonstrated applications cover the major interests in neuroanatomical research, including (1) mapping the distribution of cell types (Chapter 4.1), (2) reconstuction of the neural activity profiles by IEGs (Chapter 4.3) and (3) identification of the brain-wide circuitry using rabies virus (Chapter 4.5). Together, I propose a community-driven whole-brain mapping scheme built around CUBIC-Cloud.

[3]The PubMed (https://pubmed.ncbi.nlm.nih.gov) search result by the following query command: `(tissue clearing[Title/Abstract]) AND (("2019/01/01"[Date - Publication] : "2020/12/04"[Date - Publication]))`

Chapter 2

Materials and Methods

2.1 Sample preparation and data collection

2.1.1 Experimental animals

Wild-type C57BL/6N mice were purchased from CLEA Japan Inc. or Japan SLC Inc and housed in Ueda laboratory's mouse facility. *App*$^{\text{NL-G-F/NL-G-F}}$ mice were provided by RIKEN BioResource Research Center (RBRC No. RBRC06344) (Saito et al., 2014). Kiss1-Cre mouse was purchased from The Jackson Laboratory (Kiss1-tm1.1(Cre/EGFP)Stei/J, stock no. 017701), and housed and maintained by Touhara laboratory at the University of Tokyo. All experimental procedures and housing conditions were approved by the Animal Care and Use Committee of The University of Tokyo.

2.1.2 Tissue clearing and staining

To collect brain tissue from the mouse, animals were anesthetized by an overdose of pentobarbital ($>$ 100 mg/kg), then transcardially perfused with 10 mL of phosphate-buffered saline (PBS) and 20 mL of 4% paraformaldehyde (PFA). The dissected brain was post-fixed in 4% PFA overnight at 4 °C and stored in PBS until use in the experiments.

To clear brain tissues, I followed the second-generation CUBIC protocol (Tainaka et al., 2018). In addition, tissue staining was performed following CUBIC-HV method (Susaki et al., 2020). For the completeness, here I describe the step-by-step protocols of the CUBIC and CUBIC-HV.

PFA-fixed brain was first immersed in 50%-diluted CUBIC-L solution (10% (wt/wt) *N*-butyldiethanolamine, 10% (wt/wt) Triton X-100 in water) for overnight at 25 °C under gentle shaking. The brain was then immersed in 100% CUBIC-L solution for two to three days at 37 °C under gentle shaking. After PBS wash, the brain was placed in the nuclear staining solution. In the nuclear staining solution, nuclear staining dye was diluted in nuclear staining buffer (5% (wt/wt) Quadrol, 10% (wt/wt) Triton X-100, 10% (wt/wt) urea, 500 mM NaCl in water; Urea was omitted in some of the experiments). Depending on the nuclear staining dye, the following condition was used:

- SYTOX-G (Thermo Fisher, #S7020): 1/2500 dilution from the stock, incubate for 5 days
- BOBO-1 (Thermo Fisher, #B3582): 1/400 dilution from the stock, incubate for 5 days
- RedDot2 (biotium, #40061): 1/150 dilution from the stock, incubate for 3 days

If immunostaining step was not required, the brain was washed with PBS and subsequently processed by RI matching solution. Otherwise, the staining was performed with the following procedure. Nuclear stained brain was first washed with 10mM HEPES solution (25 °C under gentle shake). In some of the antibodies, to enhance the permeability of the tissue, hyaluronidase treatment was optionally applied with the following procedure (see Table 2.1). First, the brain was immersed in a solution containing 10 mM CAPSO and 150 mM NaCl (37 °C under gentle shake for 2 hours). Then, the brain was immersed in hyaluoridase solution (20 mg/mL hyaluronidase, 10 mM CAPSO, 150 mM NaCl in water) for 24 hours at 37 °C under gentle shake. Then, the brain was washed by the buffer (200 mM NaCl, 0.1% (vol/vol) Triton X-100, 5% (vol/vol) MeOH in 50 mM carbonate-bicarbonate buffer adjusted to pH = 10.0). After the optional hyaluronidase treatment, the brain was immersed in the HEPES-TSC buffer (10 mM HEPES, 10% (vol/vol) TritonX-100, 0.2 M NaCl, 0.5% (wt/vol) casein) for more than 1.5 hours under gentle shake at 32 or 37 °C. The HEPES-TSC buffer was occasionally supplied with 2.5% (wt/wt) Quadrol and 0.5M urea (see Table 2.1). The primary antibody was pre-mixed with secondary antibody in the HEPES-TSC buffer for 1.5 hours at 37 °C under gentle shake. Then, the brain was immersed in staining buffer. Incubation temperature and duration varied depending on the antibody (see Table 2.1). After the primary incubation, the brain was optionally incubated at 4 °C for 1 to 5 days (see Table2.1).

Lastly, the brain was processed by RI matching solution. The brain was first immersed in 50%-diluted CUBIC-R+ solution (45% (wt/wt) antipyrine, 30% (wt/wt) Nicotinamide, 0.5% (vol/vol) N-butyldiethanolamine) for overnight at 25 °C under gentle shake. On the next day, the brain was moved in 100% CUBIC-R+ solution and incubated for 3 days at 25 ° under gentle shake.

After RI matching, the brain was ready for imaging. To rigidly mount the sample on the microscope, the brain was embedded in a transparent agarose gel, following the method described by Matsumoto et al., 2019.

2.1.3 Imaging with light-sheet fluorescence microscopy

To image cleared brains, two different LSFM was used in the present study, depending on the experiment.

The first LSFM device (named GEMINI system) was custom-built by Olympus Inc. (for details please also see (Susaki et al., 2020)). GEMINI was equipped with 488, 532, 594 and 642 nm diode or DPSS lasers (SOLE-6, Omicron). The laser sheet was generated by a cylindrical lens and the sheet thickness was adjustable by a mechanical slit. For detection, the microscope was equipped with 0.63X macro-zoom objective lens (MVPLAPO 0.63X, Olympus) and 0.63-6.3X variable zoom optics (MVX-ZB10, Olympus), giving 0.4X to 4X total magnification. After passing a suitable fluorescence filter, the fluorescence signal was captured by sCMOS camera (Zyla 5.5, Andor). To achieve homogeneous light-sheet thickness throughout the field of view, GEMINI was equipped with tiled-light sheet mechanisms (see Chapter 1.2.2). In this setup, the sheet thickness was approximately 10 µm and the rectangular strip width was 1500 µm, which required 6 image strips to cover the entire brain.

The second LSFM device (named RapidScope) was developed and maintained by the imaging core of the International Research Center for Neurointelligence (IRCN) at the University of Tokyo. RapidScope was equipped with 488, 532, 594 and 642 nm diode or DPSS lasers (OBIS, Coherent). The laser sheet was generated by DSLM mechanism (see Chapter 1.2.2). In addition, axially-swept light sheet mechanism was implemented using electrically-driven liquid lenses (see Chapter 1.2.2). This

TABLE 2.1: Antibody staining conditions

Primary antibody	Dilution	Secondary antibody	Hyaluronidase	Additive	Reaction temperature	Reaction time	4 °C reaction
Parvalbumin (Swant, #PV295)	1/50	Anti mouse IgG1 Fab with Cy3 (Jackson ImmunoResearch, #115-167-185)	(+)	2.5% Quadrol	25 °C	20 days	(+)
Somatostatin (Millipore, #MAB354)	1/10	Anti mouse IgG2a with AlexaFluor 594 (Jackson ImmunoResearch, #112-587-008)	(+)	2.5% Quadrol	25 °C	20 days	(+)
Choline acetyltransferase (abcam, #ab178850)	1/200	Anti Rabbit IgG with AlexaFluor 594 (Jackson ImmunoResearch, #111-587-008)	(+)	2.5% Quadrol	25 °C	10 days	(-)
Tyrosine hydroxylase (Santa Cruz, #sc-25269)	1/20	Anti-mouse IgG2a with AlexaFluor 594 (ackson ImmunoResearch, #115-587-186)	(+)	2.5% Quadrol + 0.5M urea	37 °C	15 days	(-)
Ionized calcium-binding adapter molecule 1 (Wako #013-26471)	1/50	NA (directly conjugated with a red dye)	(+)	(-)	32 °C	10 days	(+)
c-Fos (CST #2250S)	1/75	Anti-Rabbit IgG with AlexaFluor594 (Jackson ImmunoResearch, #111-587-008)	(-)	(-)	32 °C	10 days	(+)
6E10 (Biolegend, #93049)	1/100	anti-mouse IgG1 with Alexa 594 (Jackson ImmunoResearch, #115-587-185)	(+)	2.5% Quadrol	25 °C	14 days	(-)

mechanism allowed homogeneous resolution in XYZ. For detection, the microscope was equipped with 0.63X macro-zoom objective lens (MVPLAPO 0.63X, Olympus) and 0.63-6.3X variable zoom optics (MVX-ZB10, Olympus), giving 0.4X to 4X total magnification. After passing a suitable fluorescence filter, the fluorescence signal was captured by sCMOS camera (pco.edge 5.5, PCO).

For each dye/FP, the following laser and fluorescence filter pair was used: AlexaFluor 594 [Ex: 594 nm, Em: 641/75 nm bandpass (FF02-641/75-32, Semrock)], Cy3 [Ex: 532 nm, Em: 585/40 nm bandpass (FF01-585/40-32, Semrock)], SYTOX-G, BOBO-1 and GFP [Ex: 488 nm, Em: 520/40 nm bandpass (FF01-520/44-32, Semrock)], RedDot2 [Ex: 642nm, Em: 708/75 nm bandpass (FF01-708/75-32, Semrock)], mCherry [Ex: 594 nm, Em: 628/32 nm bandpass (FF01-628/32-32, Semrock)].

2.1.4 Imaging the fluorescent bead embedded in cleared tissue

An essential prerequisite for whole-brain analysis using tissue clearing is that the image quality is identical throughout the entire brain, without attenuation or blurring. To validate this condition, I imaged the fluorescent beads embedded in the tissue with LSFM and measured the spot profile.

1.0 µm-diameter green-yellow fluorescent beads (Thermo Fisher, #F8765) were diluted in PBS (the final bead concentration was 0.9×10^7 particles/ml). This bead-mixed PBS solution was perfused in mice, prior to the PFA perfusion. Because the bead surface was modified with amine, PFA was able to cross-link and fix the beads within the tissue. After tissue clearing with CUBIC, the whole brain was imaged using the macro-zoom LSFM (GEMINI) with XYZ voxel resolution of 6.45x6.45x7.0 µm (Figure 2.1 **A**). Then, single and well-isolated bead particles were manually picked up ($n > 15$ for each brain region). Subsequently, the mean spot profiles were computed and fitted with Gaussian. The fitted sigma values from six regions were all within 4.4 to 4.9 µm (lateral) and 5.3 to 6.7 µm (axial) (Figure 2.1 **B**), validating homogeneous image quality throughout the entire brain. Given the digital sampling frequency (6.5 µm) of the microscope used, this result was nearly the ideal PSF.

2.1.5 Rabies virus production and injection

Note: RV and AAV production and injection was done by Dr. Murata and Dr. Miyamichi from Touhara laboratory at the University of Tokyo. For the completeness, I describe the virus production and injection protocols here.

The following AAV vectors were generated de novo by PENN vector core using the corresponding plasmids. AAV serotype 9 CAG-FLEx-TCb (1.5×10^{13} gp/ml) was made using the plasmid described previously (Miyamichi et al., 2013). Here TCb stands for TVA-mCherry expression cassette optimized to increase mCherry brightness. To generate AAV serotype 9 CAG-FLEx-oG (4.5×10^{13} gp/ml), engineered and optimized glycoprotein (oG) (Kim et al., 2016) sequence was ligated to pAAV-FLEX sequence from pAAV-FLEX-GFP (Addgene).

Preparation of rabies virus was conducted by using the RVΔG-GFP, B7GG and BHK-EnvA cells as previously described (Osakada and Callaway, 2013). The EnvA-pseudotyped RVΔG-GFP+EnvA titer was estimated to be 1.0×10^9 infectious particles/ml based on serial dilutions of the virus stock followed by infection of the HEK293-TVA800 cell line.

For trans-synaptic tracing using rabies virus, about 20 nL of mixture of AAV9 CAG-FLEx-TCb and CAG-FLEx-oG (diluted to 1.5×10^{12} gp/ml each) was injected

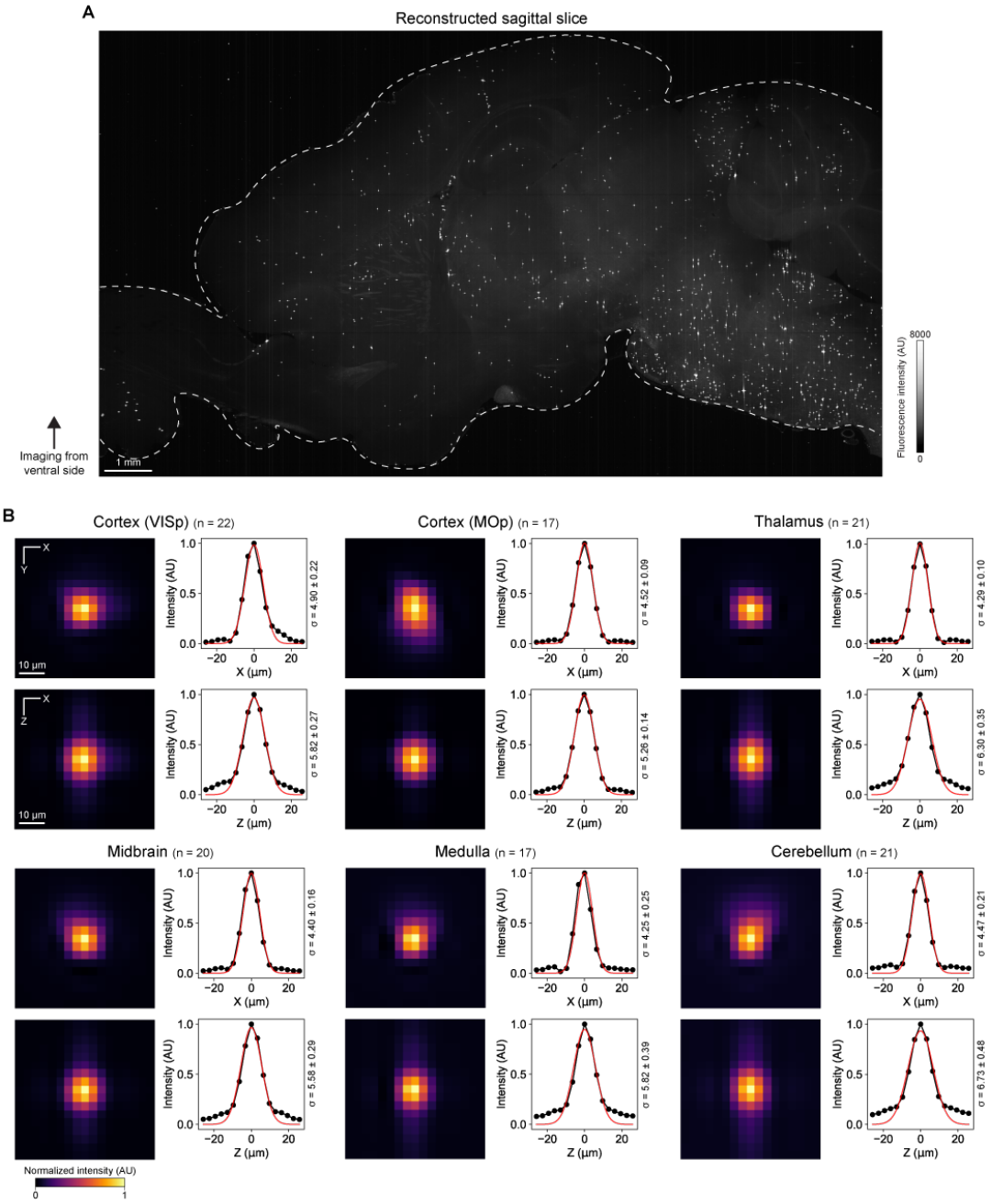

FIGURE 2.1: Measurement of the fluorescent bead embedded in the cleared tissue

A. Representative sagittal slice (virtually reconstructed from horizontal-major 3D image stack) showing the fluorescent beads embedded in tissue. **B**. Bead spot profiles measured in six brain regions. Lateral and axial profiles were fitted with Gaussian, respectively, and the fitted curves are shown along with the raw data points. The fitted sigma values (with 95% confidence interval) of the Gaussian are also shown. The number of particles used to average the spot profile are shown in the graph.

into the ARH of Kiss1-Cre mice. The first AAV transduced a TVA receptor fused with mCherry for EnvA. The second AAV transduced RV glycoprotein (oG) playing a predominant role in the trans-synaptic transport of RV. The injection coordinate was P1.1, L0.2, V5.9 (distance in mm from the Bregma for the posterior [P], and lateral left [L] positions and from the brain surface for the ventral [V] position). Three

weeks later, 30 nL of Rabies ΔG-GFP+EnvA was injected into the same brain region to initiate trans-synaptic tracing. Because there is no cognate receptor for EnvA in the mouse brain, RVΔG+EnvA only infects TVA-expressing cells. oG expression from the second AAV complements the RVΔG, allowing retrograde monosynaptic tracing from Cre-expressing cells. Seven days later, brains were sampled for CUBIC treatment.

2.2 Image analysis methods used in CUBIC-Cloud

2.2.1 Cell segmentation and detection

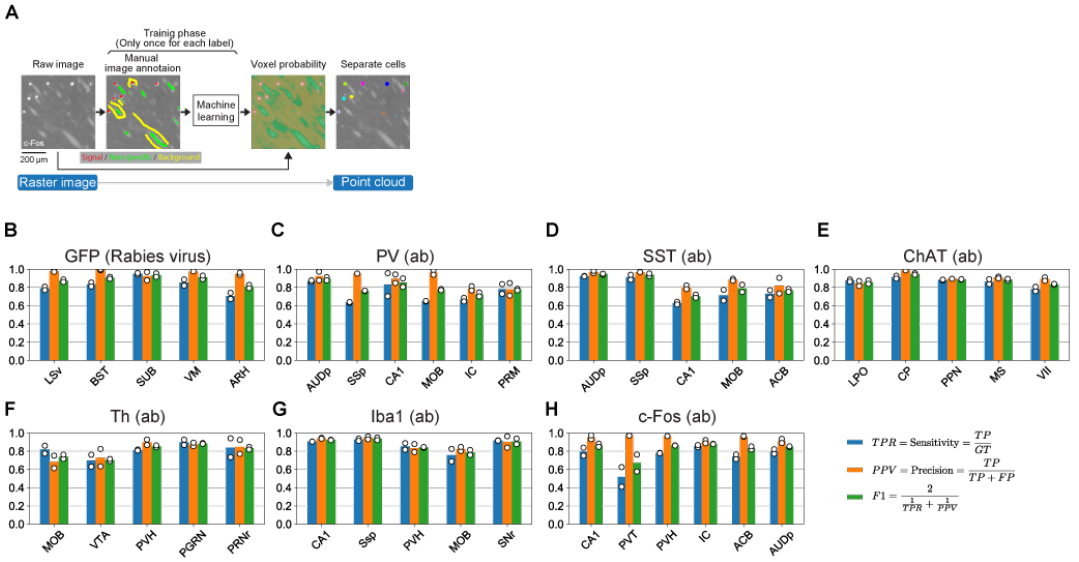

FIGURE 2.2: Cell counting method and its accuracy

A. The cell detection workflow overview (also see Chapter 2.2.1). **B-J**. Accuracy evaluation of the cell counting algorithm. True positive rate (TPR), positive predictive value (PPV) and F1 score were evaluated for seven label types in more than five brain regions. Ground truth was prepared by two independent annotators.

The cell detection from 3D brain image data was performed with the following procedure (Figure 2.2 **A**). Because the raw 3D brain image data contained noises and high-intensity background which varied depending on the brain region, simple image filtering approaches did not work sufficiently well to isolate labelled cells. Thus, I used ilastik (Sommer et al., 2011; Berg et al., 2019) software. ilastik accepts hand-crafted image features, and uses random forest algorithm to classify the 3D voxels into distinct classes. ilastik was chosen over other cell segmentation software/library, because (1) ilastik can be trained with sparse annotation data set, which reduced the cost of having to prepare dense annotation and (2) the code was highly parallelized and scalable to many-core servers. To train the classifier, manual annotation images were prepared. In this study, three classes were defined, which were (1) signals of interest, i.e., cells labelled by FPs or antibodies (2) bright but false signals, such as non-specific binding of antibodies to vascular structures or neurites extending from cell bodies and (3) background (i.e. void space). Typically, 5,000 to 10,000 voxels were annotated as class 1 per one dataset. To increase the robustness,

at least two brains with identical labelling conditions were annotated. Image annotation was performed using ITK-SNAP software (Yushkevich et al., 2006). Next, following the ilastik workflow, image feature descriptors were selected, which was fed into the random forest algorithm. For PV, Sst, ChAT, Th, Iba1, c-Fos, RV-GFP and AAV-mCherry images, the selected descriptors were Gaussian (σ = 0.3, 0.7 voxel), Gaussian gradient magnitude (σ = 0.7 voxel), difference of Gaussian (σ = 0.7, 1.0, 1.6, 3.5 voxel) and Hessian of Gaussian eigenvalues (σ = 1.0, 1.6, 3.5 voxel). For the analysis of 6E10-labeled Aβ images, difference of Gaussian (σ = 5.0 voxel) and Hessian of Gaussian eigenvalues (σ = 5.0 voxel) were additionally included, so that the larger spatial context was taken into account. Then, hyperparameters in random forest algorithm was automatically optimized by ilastik.

By applying the voxel classifier trained above to each brain image, a probability image was produced, where the value of each voxel represents the probability of that voxel being class 1. The probability value was given in the range [0, 1]. Using this probability image, a custom Python program was used to isolate individual cells in the following way. First, the probability threshold, $P_{\mathrm{th}} = 0.7$, was applied to make a binarized image. Then, connected voxels were searched and merged together, to find individual objects. If the identified object volume was larger than a threshold, V_{th}, it was sent to the object separation routine. The object separation routine simply finds local maxima with an exclusion distance $r_{\mathrm{excl}} = V_{\mathrm{th}}^{1/3}$. V_{th} was heuristically determined to be $V_{\mathrm{th}} = 4^3 = 64$ for PV, Sst, c-Fos, Iba1, Rabeis-GFP and AAV-mCherry, and $V_{\mathrm{th}} = 5^3 = 125$ for ChAT and Th. For Aβ plaque segmentation, $V_{\mathrm{th}} = \infty$ was used. In this way, all of the single cells from the 3D brain image is isolated. The final output was written in a comma-separated values (CSV) format, which recorded the XYZ position, the mean and maximum fluorescent intensity over the object, and the volume of the object.

Accuracy of the above explained cell detection procedure was extensively evaluated by comparing automated counting results with manual cell counting (Figure 2.2 **B**-**H**). Manual cell counting was performed by cropping a small cubic image volume (50 or 75 or 100 voxels, depending on the cell density) from brain images which were not used in machine learning training. Well-trained human annotators ($n = 2$) independently marked all of the cells present in the image and typically yielded 100-200 marked cells. Image annotation was performed using ITK-SNAP software (Yushkevich et al., 2006). Cells annotated by both human and algorithm were regarded as true positives. Cells annotated by human but not by algorithm was regarded as false negative. Cells annotated by algorithm but not by human were regarded as false positives. Then, true positive rate (TPR, also called sensitivity) and positive predictive value (PPV, also called precision) were evaluated for the ground truth annotation prepared by each human annotator. To quantify the overall performance, F1 score, defined as $F1 = 2 * (PPV \times TPR)/(PPV + TPR)$, was also evaluated. For most of the label types and brain regions, our cell detection algorithm robustly demonstrated good F1 scores, with average score being 0.80 (PV), 0.83 (Sst), 0.88 (ChAT), 0.80 (Th), 0.88 (Iba1), 0.83 (c-Fos) and 0.89 (GFP).

2.2.2 Brain registration

CUBIC-Cloud uses the symmetric image normalization (SyN) algorithm implemented in ANTs library (Avants et al., 2008) to run registration between CUBIC-Atlas ("fixed" brain) and individual brain sample ("moving" brain). First, nuclear staining image of

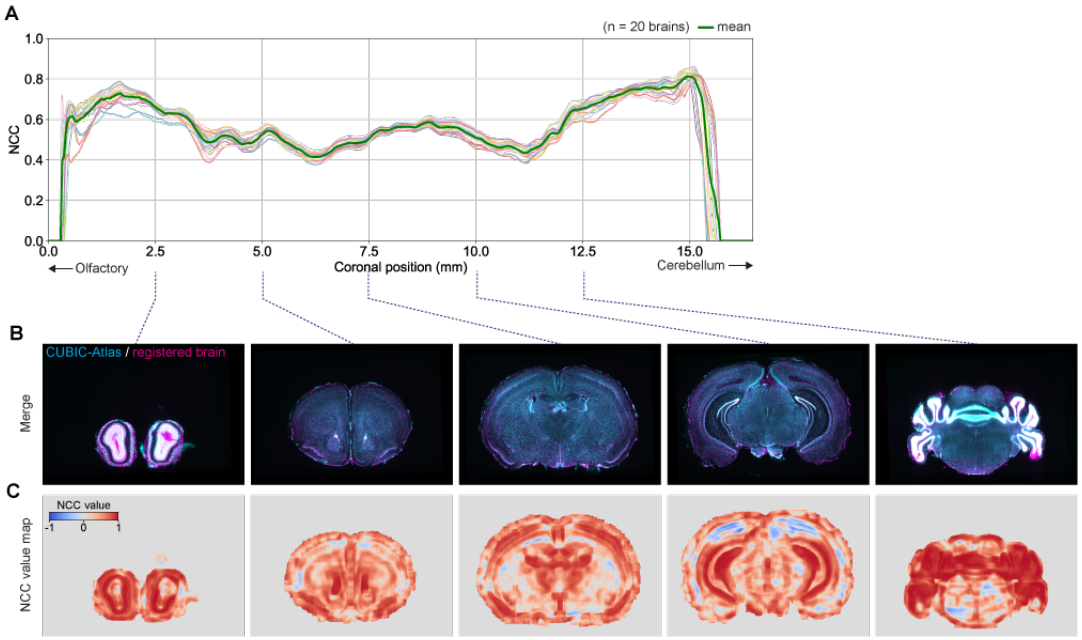

FIGURE 2.3: Brain registration method used in CUBIC-Cloud

A. Normalized cross-correlation (NCC) value between two brains after registration. Mean NCC values of each coronal slice are plotted. 20 brains from different mice were independently registered onto CUBIC-Atlas. Individual profiles (thin lines with light colors) as well as the mean (thick green line) are shown. **B**. Representative brain registration result. CUBIC-Atlas (cyan) and registered brain (magenta) are overlaid. **C**. Voxel-wise NCC value map computed for the images shown in **B**.

the "moving" image was downscaled to a voxel size of 50x50x50 µm. Nuclear staining image of CUBIC-Atlas was downscaled to a voxel size of 80x80x80 µm. Considering the sample's physical expansion by clearing treatment (2.2X for CUBIC-Atlas and 1.5X for CUBIC-R+ treated brains), this downscaling operation resulted in an effective voxel size of about 35x35x35 µm in both images. The registration first computed affine transformation to coarsely align the orientation and size, using mutual information as the optimizer metric. Subsequently, non-linear warping was computed by SyN algorithm, which optimized the warp field by maximizing the normalized cross-correlation (NCC) between the two images under diffeomorphic regularization (Avants et al., 2008). Given image $I(\mathbf{x})$ and image $J(\mathbf{x})$, the NCC value between I and J at the voxel position $\mathbf{x}$ is given by

$$NCC(I, J, \mathbf{x}) = \frac{\langle \bar{I}, \bar{J} \rangle}{|\bar{I}|\,|\bar{J}|} \tag{2.1}$$

where $\langle A, B \rangle$ represents the inner product taken over a local window with radius R centered at position $\mathbf{x}$. $|A|$ is the L2 norm of the vector computed over a local window with radius R. Here, $\bar{I}(\mathbf{x}) = I(\mathbf{x}) - \mu_I(\mathbf{x})$ means the subtraction of the local mean, where local mean $\mu_I(\mathbf{x})$ is computed over a local window with radius R centered at position $\mathbf{x}$. $R = 4$ (voxels) was used in all analysis.

The representative registration result is visualized in Figure 2.3 **B**, along with the NCC value heatmap (Figure 2.3 **C**). To show the reproducibility of the registration, 20 individual brains were mapped onto CUBIC-Atlas with identical registration parameters. The mean NCC value of each coronal planes were computed and plotted

(Figure 2.3 **A**). 20 independent curves overlap with each other, meaning that the optimization attempt by the registration reached saturation. NCC value tends to show higher value in the olfactory area and cerebellum, due to the presence of more distinct structural features.

2.3 Details on data analysis

2.3.1 Whole-brain analysis of RV-injected brains

After AAV and RV injection, Kiss1-Cre mice (13- or 14-week-old at the time of brain sampling) were cleared, stained, imaged and analyzed as described in the corresponding sections in Chapter 2. Brains were stained with nuclear staining dye (RedDot2, Biotium, #40061).

As a negative control experiment, AAV and RV injection was performed using BALB/c wild-type mice brain ($n = 3$). After clearing, the whole-brain image was obtained by LSFM. No GFP or mCherry signals were observed by manual inspection, confirming the absence of Cre-indepentent leakage of AAV vectors and specificity of the virus delivery (data not shown).

Starter cells were searched by identifying dual positive (mCherry+ and GFP+) cells. For each channel, cell counting was independently performed, and the center of the mass of the detected cell was obtained. For each mCherry+ cells, if a GFP+ cell was present within a distance of 24 µm, the cell was counted as starter[1]. Occasionally, slight voxel shift (typically no more than 4 voxels) occurred between GFP and mCherry channels, which was presumably caused by slight misalignment between the 488 nm and 594 nm illumination laser or drift of the specimen during scanning. To correct this, small 3D volumes with distinct features (typically (X,Y,Z) = (50,50,20) voxels, $n = 4$ or $n = 3$) from mCherry and GFP channels were cropped, and the voxel shift was computed by registering two images using ANTs, where transformation was restricted to only translation. Then, the cell coordinates were corrected by the mean of the computed shift.

To carry out statistical analysis of input cell numbers between male and female brains, I used the normalized cell count, $n_{\mathrm{norm},i}$, where i represents the ID of the brain region. Denoting the raw cell number of each brain region as $n_{\mathrm{raw},i}$, $n_{\mathrm{norm},i}$ is simply expressed as $n_{\mathrm{norm},i} = n_{\mathrm{raw},i}/(\sum_i n_{\mathrm{raw},i})$.

2.4 Implementation details of CUBIC-Cloud

2.4.1 Cloud infrastructure built upon serverless architecture

The schematic illustration of the cloud architecture is shown in Figure 2.4. CUBIC-Cloud is constructed using the serverless architecture and deployed on the Amazon Web Service (AWS). When user accesses the web site, the static site content is distributed by CloudFront. CloudFront is responsible for caching, managing secure connection through SSL/TLS and web application firewall (WAF) to reject malicious access. Then, static web contents are fetched from S3 bucket and returned to the user. User authentication is handled by Cognito. Once authenticated, users can access the protected API endpoints securely using json web token (JWT). All REST API requests are routed by API Gateway. API Gateway forwards most of the API

[1] Note that because the CUBIC-cleared tissue was expanded by a factor of ∼1.5, this was roughly 16 µm in untreated tissue.

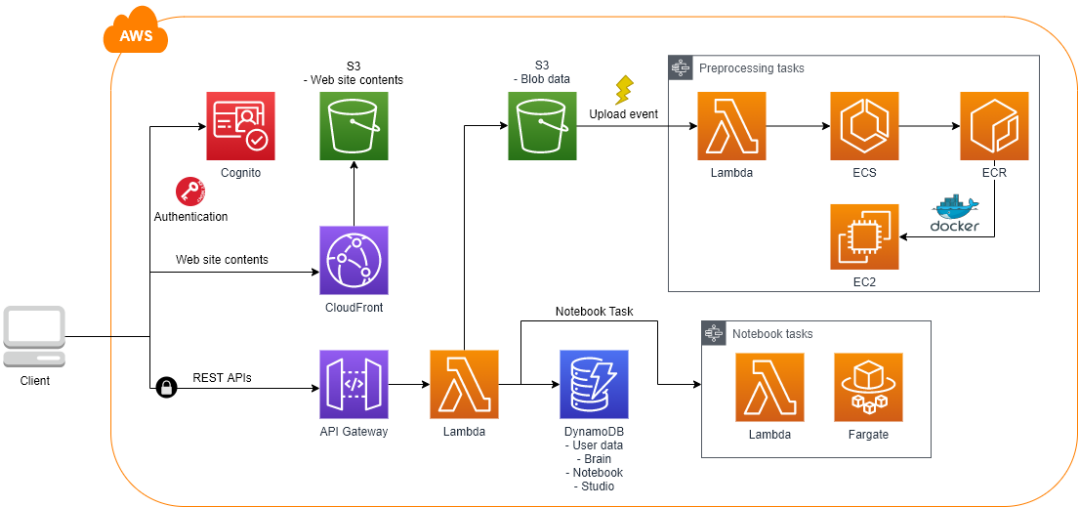

FIGURE 2.4: Architecture of CUBIC-Cloud

The diagram shows the schematic architecture of CUBIC-Cloud (see Chapter 2.4.1 for details).

request to Lambda. Lambda is the most essential serverless computing unit in AWS, which can execute a low-workload tasks which would complete in milliseconds to a few seconds. In CUBIC-Cloud, computation that do not require CPU power such as read and write operation on the database, are all handled by Lambda. Lambda handlers have access to various back-end resources, including the databases and the data buckets. As a serverless non-SQL database, DynamoDB, was used to store information on users, brains, notebooks, and studios. Large data files (such as images and csv tables) were stored in S3.

Once a user uploads the brain data, the upload completion event is triggered from S3 to start a "preprocessing" task in the ECS cluster. Preprocessing includes brain registration, transformation, and data conversion. ECS automatically launches a new EC2 instance, pulls the Docker container from ECR registry, and initiate a new task. The task execution is orchestrated by StepFunctions. Notebook tasks (i.e. generating plots) are similarly orchestrated by StepFunctions, except that the runtime is either Lambda or Fargate, depending on the required memory size of the task.

The cloud application stack was written with AWS's Cloud Development Kit (CDK) framework for Python (`https://github.com/aws/aws-cdk`). The API handlers executed by Lambda are written in Python and boto3 library (`https://github.com/boto/boto3`) was used to manipulate other AWS resources, such as DynamoDB and S3. The graphical user interfaces (GUIs) on the web browser was created using Vue.js framework (`https://github.com/vuejs/vue`).

2.4.2 Implementation of the 3D brain viewer

CUBIC-Cloud offers a point-cloud based interactive 3D brain viewer, a feature called studio. The viewer is written in JavaScript, and runs in the standard web browsers, including Google Chrome and Firefox. It uses WebGL (`https://www.khronos.org/webgl/`) for hardware accelerated 3D rendering. Three.js library (`https://github.com/mrdoob/three.js/`) was used to write code for graphics rendering. The core of the point cloud rendering engine was adopted from the open-source

project, Potree (Schuetz, 2016). Following the implementation of Potree, the raw point cloud data is converted into a chunked format, where the whole point cloud was divided and stored in a multi-resolution hierarchical structure (octree structure). This octree-formatted data was stored in the server with a UNIX-based file paths, so that the chunk can be fetched from the client through standard GET HTTP request. The octree-formatted point cloud data was automatically generated as a part of the preprocessing task in the CUBIC-Cloud server. On the client side, point cloud data is adaptively queried in response to user's viewpoint, in which a portion of the points near the viewer's camera was loaded with high priority.

Each point can be attached with several attributes, including the brain region ID and fluorescent intensity. Points may be colored or filtered using these attributes. For example, points may be given gradient colors based on their fluorescence intensity values, or regions in specific brain regions may be selectively shown using the point's region ID.

Chapter 3

Results 1: Construction of CUBIC-Cloud

In chapter 1.4, I formulated the software challenges that will be required in effective integration of the whole-brain datasets collected by neuroscience research community. Based on this motivation, I implemented a cloud-based application for whole mouse brain analysis, which I named CUBIC-Cloud. The CUBIC-Cloud web site is hosted at `https://cubic-cloud.com`.

In summary, CUBIC-Cloud offers the following functionalities:

- Uploading the brain data to the cloud, and automatically register it to the reference brain.
- Constructing the user's own brain repository in the cloud.
- Run 3D interactive visualization of the brains using "studio".
- Run quantification of the brains using "notebook".
- Sharing and publishing of the brain data, studio and notebook results.

Below, I will explain the design considerations and rationals of the cloud system design (chapter 3.1). Then I describe the front-facing software functionalities (chapter 3.2) and the back-end implementation details (chapter 3.3). A step-by-step user guide of CUBIC-Cloud is provided in Appendix B.

Contribution statement: The majority of the CUBIC-Cloud application stack was designed and implemented by the author. The cloud server development was assisted by R. G. Yamada and S. Horiguchi. The development of the web graphical user interface was assisted by Tecotec Inc.

3.1 CUBIC-Cloud: Considerations and rationals

3.1.1 Choice of the reference brain

Mapping of the submitted brains to the reference brain is the most important step in CUBIC-Cloud. Therefore, the choice of the reference brain needs to be carefully considered.

In this study, the following three mouse brain atlas were considered as candidates. The first two candidates were provided by Allen Mouse Brain Common Coordinate Framework version 3 (CCFv3) (Wang et al., 2020). In CCFv3, two variant formats are provided. The Nissl staining atlas (Figure 3.1 **B**) is a brain structure labeled with Nissl staining, constructed from the serial sectioning imaging method.

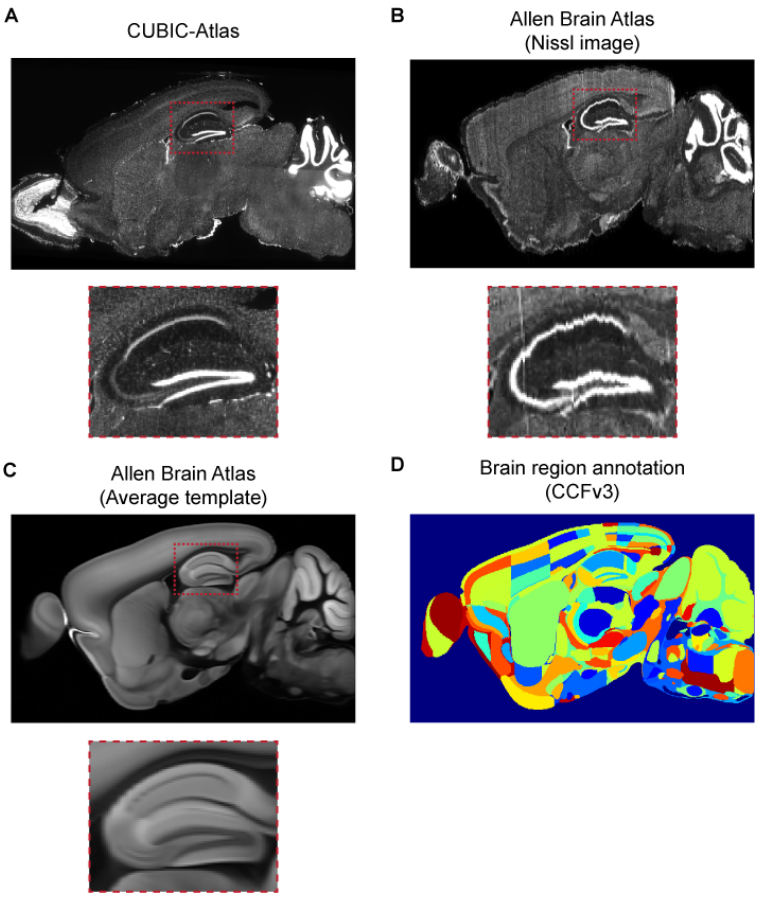

FIGURE 3.1: Comparison of the mouse brain atlas

A-C. Representative sagittal slice from CUBIC-Atlas (**A**), Allen Brain Atlas Nissl staining image (**B**) and Allen Brain Atlas Averaged brain (**C**). The enlarged view of the hippocampus is shown in the lower panel. **D**. The brain area annotation from CCFv3. Each brain areas are colored by a unique RGB value.

The averaged template (Figure 3.1 **C**) is a brain structure constructed from the average of 1600 adult mouse brains, scanned with serial two photon tomography visualizing the autofluorescence of the tissue. In CCFv3, brain region annotation is provided, which is aligned with Nissl image and average template (Figure 3.1 **D**). The third candidate is CUBIC-Atlas (Figure 3.1 **A**) (Murakami et al., 2018). CUBIC-Atlas was constructed by tissue clearing and high-resolution LSFM imaging. The brain region annotation was imported from CCFv3 via semi-automatic registration. Thus, effectively the same brain area annotation from CCFv3 are defied on CUBIC-Atlas.

In the present study, CUBIC-Atlas was used as the reference brain based on the following considerations.

- Nissl image from CCFv3 contains artifacts arising from the serial sectioning (as is evident in the zoom-in view of the hippocampus in Figure 3.1 **B**). Due to the concern that this artifact cause unexpected errors in registration, Nissl image from CCFv3 was rejected.
- Average template of CCFv3 records the autofluorescence of the tissue. On the other hand, CUBIC-Atlas records the nuclear staining. Some brain structure,

such as the layers of the cortex and the hippocampus, are more clearly visible in nuclear staining images. For this reason, nuclear staining was evaluated to be a better structural marker.

- CUBIC-Atlas offers information about the absolute number of cells in each anatomical regions, which is useful when normalization by the total cell number is necessary. This information is not present in CCFv3.

Nonetheless, it is important to point out that the mouse brain atlas is constantly renewed over the years, so there is a possibility to a newer reference brain. Thus, in the current implementation, CUBIC-Cloud stores the atlas version information as meta-information for each of the brain.

3.1.2 Data format

The next important consideration is defining the data format accepted in the cloud system. As discussed in the human neuroimaging database (chapter 1.4.4), a separation of raw image database and the curated database would be more reasonable approach, since the they would require different mission and software design. In this study, CUBIC-Cloud was designed to collect curated whole mouse brain data. In particular, CUBIC-Cloud accepts the whole-brain mapping data represented in point cloud format. This is a representation where each cell segmented from the raw image is abstracted as a single point, carrying the following biological information:

- XYZ position of the center of the mass.
- Mean and maximum fluorescent intensity over the object.
- The volume of the object.

This design choice was made because most of the brain mapping project would be most interested in the quantification of the number and the expression level of the targeted cells. For example, in the gene expression mapping applications (Chapter 1.1.1), the goal is to quantify the number of cells expressing the gene and the amount of the gene expression of each cell, of all brain areas. Thus, point abstraction still gives essential biological information that is of interest in most of the studies, while drastically reducing the file size from raw images.

It is certainly true that the point cloud abstraction may not be sufficient to represent all biological information. Other possible input formats are vectors (e.g. to represent neural fibers) and polygons (e.g. to represent detailed morphology of the cells and synapses). Accepting these various data format is highly attractive, but considering that CUBIC-Cloud is the first experimental attempt toward data sharing, I considered that these advanced formats will be better treated in the future developments. Possible integration with other abstracted data format will be discussed in Chapter 5.2.3.

In addition to the point cloud data, users would need to submit brain structure image to run brain registration. In the current implementation, the structural information should be supplied as the whole-brain nuclear staining image.

In summary, CUBIC-Cloud requires the user to submit the following data: (1) whole-brain nuclear staining image and (2) the list of labeled cells represented as point cloud.

3.1.3 Labeling scheme of the brain data

One of the important lessons from the human neuroimaging database platforms is that the coherent and organized labeling of the brain data facilitates the re- and meta-analysis (Chapter 1.4.4). Based on this requirement, CUBIC-Cloud lets the users to supply various attributes to the brain data. The attributes currently available includes the following:

- Title of the data.
- Mouse line information (useful for transgenic animals).
- Age of the mouse.
- Label tags. Each tag has its own URL, where users can read the information on the labeling targets (for example, antibody target or fluorescent protein drivers)
- Project tags. Each tag has its own URL, where users can read the information on the experimental conditions.
- Link to the published paper.
- Free text notes to describe other information

Usually, one set of experiment would study several brains with the identical condition to evaluate the statistical significance. The "project tag" will be useful to link the brains collected under the same experiment. The "label" tag would usually indicate the protein name targeted by the experiment. Using this attribute, one would be able to search several experiments that targeted the protein of interest across the database.

Currently, to maximize the ease of use and flexibility, no ontology sets are defined to describe the above attributes. Depending on the future uses by the community, though, the mandatory use of the defined ontology may be beneficial.

3.2 CUBIC-Cloud workflow

The standard user workflow of CUBIC-Cloud is illustrated in Figure 3.2. In summary, the workflow divides into (1) tissue clearing and image collection, (2) single cell detection, and (3) uploading the data to CUBIC-Cloud, where brain registration, brain-wide quantification and visualization are performed.

3.2.1 Uploading brain data

The first step of the CUBIC-Cloud workflow is to collect the whole-brain images and upload it to the cloud. For those users that are only interested in the mining of the public datasets, this step may be skipped.

In the present study, second-generation CUBIC method was used to clear the brain tissue (Tainaka et al., 2018) (see chapter 2.1.2). The cloud system should be compatible with other clearing methods, as long as the similar level of tissue transparency and morphological preservation is ensured (more discussion on this in chapter 5.2.1). The cleared tissue was scanned with LSFM (see Chapter 2.1.3). As described in chapter 3.1.2, one of the channel was used for nuclear staining and the other channels were for the labeling of the cells targeted in the experiment.

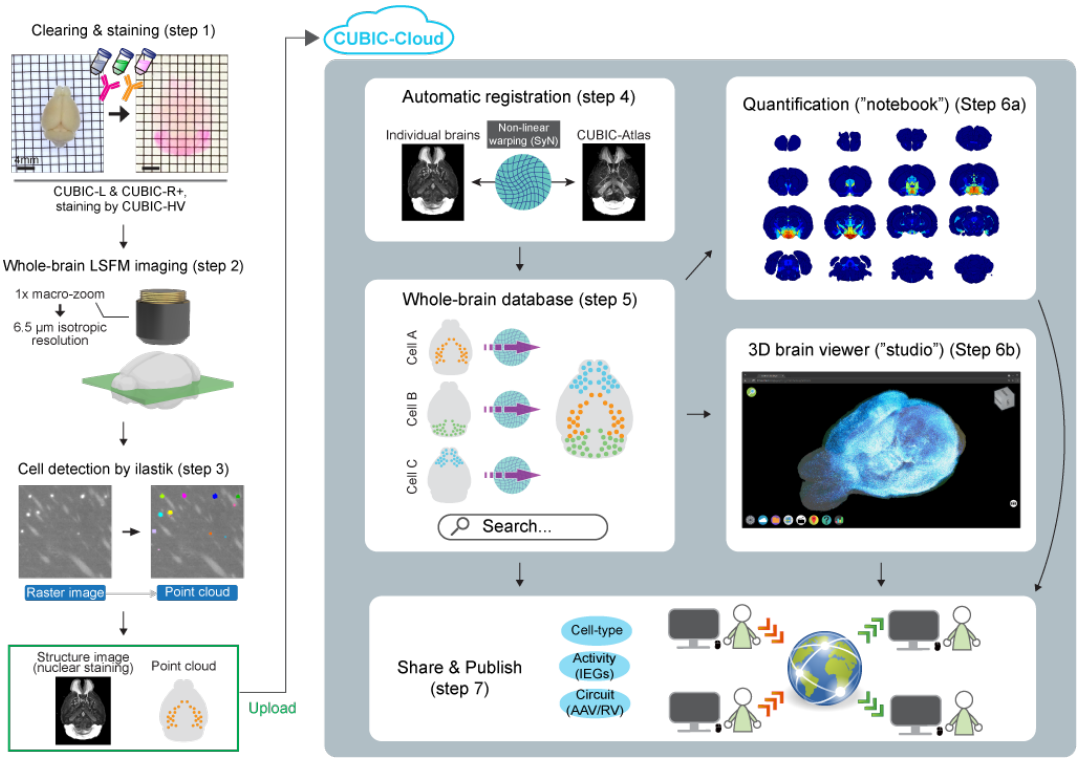

FIGURE 3.2: Overview of CUBIC-Cloud workflow

The second step of the workflow is isolation of single cells using automated object segmentation to prepare the cell data in point cloud format. In the proposed pipeline, machine learning-based cell segmentation software, ilastik (Sommer et al., 2011), was used (see chapter 2.2.1). ilastik was highly suitable for the cell segmentation of the whole-brain image data sets, because it is fast enough to handle whole-brain images and comes with a GUI to interactively train a machine learning model. However, since the cell segmentation task is highly dependent on the experiment settings, the users may use their own custom segmentation method. As long as the final output follows the CUBIC-Cloud-specified format, CUBIC-Cloud is able to process those data as well.

User would then use the GUI to upload two data files (nuclear staining image and segmented cell table) to CUBIC-Cloud. Once data upload is complete, the uploaded brain is placed in the "preprocessing" task in the cloud. The main purpose of preprocessing is running brain registration to map the submitted brain to the reference brain coordinate (for more information, see Chapter 2.2.2), along with other post-processing of the data. After the preprocessing is complete, the data is registered in the user's cloud brain repository, and is ready to be used in visualization and quantitative analysis.

3.2.2 Organizing brain repository

By uploading the brains to the cloud, users can construct their own brain repository in the cloud. All uploaded brain initially goes to the user's private storage space, so at this point the brains are not publicly visible. The list of the brains in the user's brain repository can be viewed in a table. Users would add appropriate attributes to

the brain data (as discussed in Chapter 3.1.3) to organize their brain database. User can search or filter the brains using the attributes attached to the brain.

3.2.3 Visualizing whole-brain data

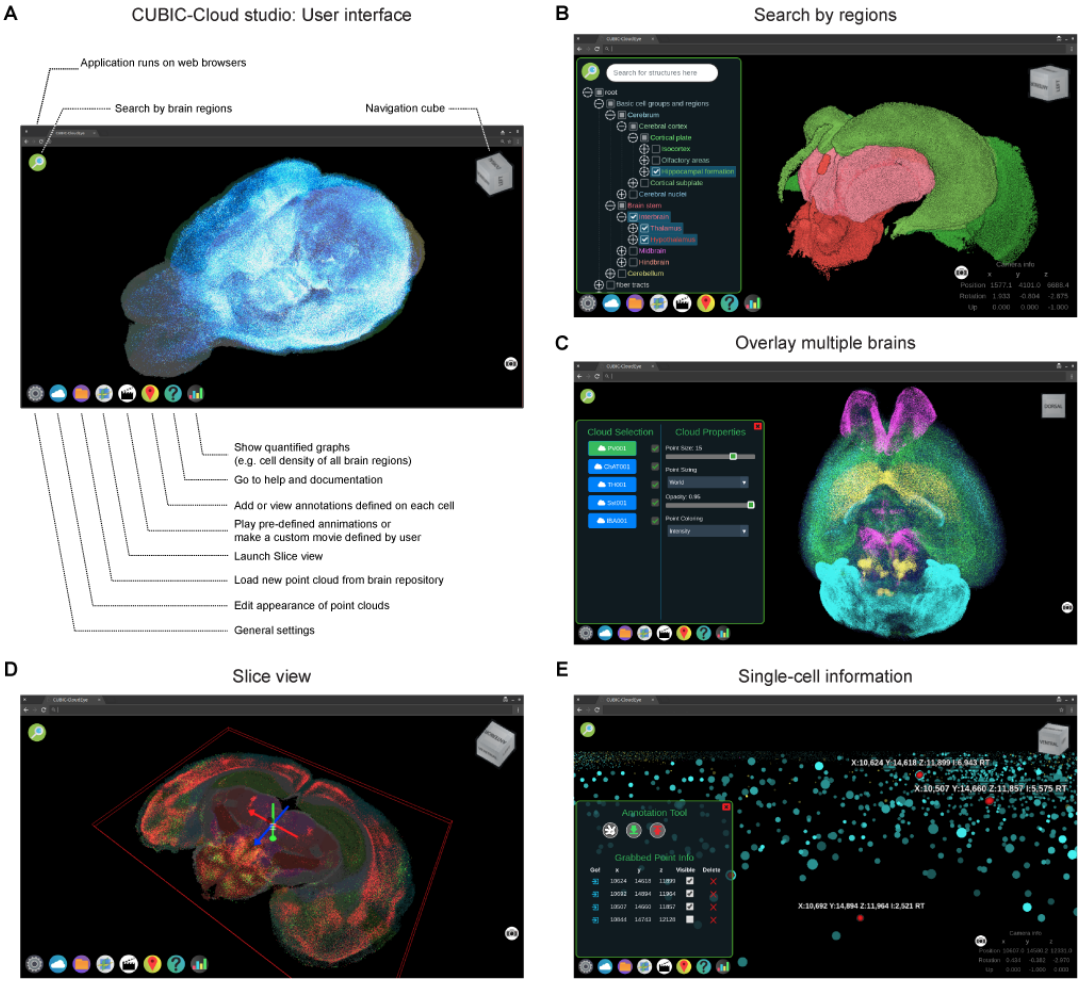

FIGURE 3.3: Studio function offered in CUBIC-Cloud

A. Overview of the user interface. **B**. User can show only the selected brain regions. **C**. Arbitrary number of brains from the database can be overlaid with user-defined colors. **D**. Section view with arbitrary angle and arbitrary thickness can be created. **E**. User can touch a single point and query the information about the cell.

Whole-brain data uploaded to CUBIC-Cloud typically contains tens of thousands to millions of single-cells scattered in 3D space. Intuitively understanding and navigating through such complex dataset is a big challenge. Modeling after GenomeBrowser (Karolchik, Hinrichs, and Kent, 2009), CUBIC-Cloud offers a web-based interactive 3D brain viewer, a feature called "studio" (Figure 3.3). The studio is a light-weight point cloud renderer that natively runs on the web browser using WebGL framework. Here, brain data is visualized as a point cloud, where each point corresponds to a single cell isolated from the raw raster data. In the server, point cloud data is stored in a hierarchical chunked format (see Chapter 2.4.2). This allows the client application to adaptively query the points based on the position of the

user's point of view. This adaptive querying mechanism allows to efficiently visualize enormous number of cells in real time even with a limited network bandwidth and graphics power. Each point carries various biological attributes such as intensity values from the raw image, object volume and the brain region IDs. The studio is able to render the points based on these attributes. For example, the renderer can assign gradient of colors based on the intensity value of the cell. Combined with maximum intensity projection (MIP) rendering, this gives effective "see-through" view of the brain, where regions with high intensity values are highlighted.

Other useful user functions include the following.

- Pan, rotate and zoom: Users can manipulate the model with a simple mouse interaction. The interface is compatible with touch screens on mobile and tablet devices.
- Region view (Figure 3.3 **B**): User can show the selected brain region and hide other regions. This feature is useful to focus on the regions deep in the brain. User would select which regions to show via GUI, where the hierachical anatomical regions are displayed as a tree.
- Overlaying multiple brains (Figure 3.3 **C**): Arbitrary number of brains can be overlaid together, each with user-defined color. This feature is possible because all brains in the repository is pre-aligned with the reference. Using this function, user would visually inspect the co-localization of the cells in two different datasets.
- Section view (Figure 3.3 **D**): User can selectively see the sections from the 3D volume. The orientation and the thickness of the section can be freely changed, either using navigation GUI or by typing numerical values in the text box.
- Movie generation: Users can record the movie in which the camera moves along the trajectory defined by the user. In addition, some preset motions are available, such as simple rotation.

3.2.4 Quantifying whole-brain data

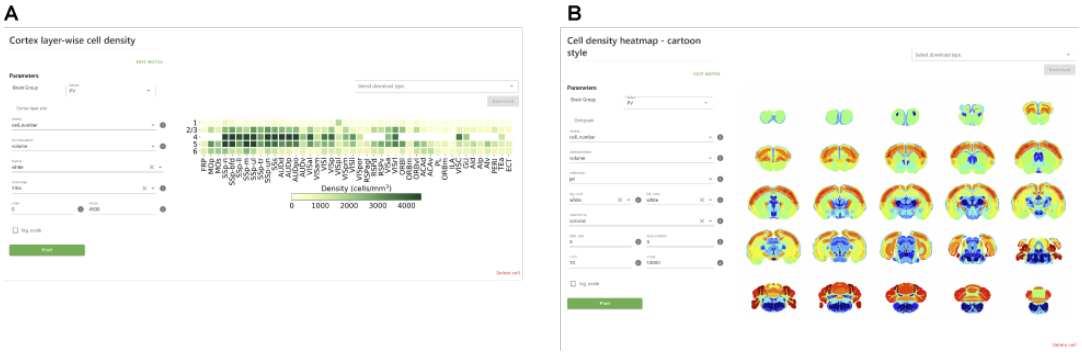

FIGURE 3.4: Representative applet GUIs offered in the notebook.

A. An applet to analyze the layer-wise density or expression levels of the isocortex. **B**. An applet to generate a cartoon-style whole-brain map of cell density or expression levels.

The studio tool described in the previous section is useful in gaining qualitative and visual understanding of the whole-brain data. To get quantitative understanding of the data, CUBIC-Cloud offers a "notebook", a feature that allows users to create various kinds of graphs with simple GUIs without writing codes (Figure 3.4). As the name signifies, the notebook allows the user to organize their analysis based on a notebook-like document and add different kind of analysis under the single document. The user first select the brains to analyze from their own brain database or from the public brain repository. Then, several pre-defined applets are provided, which allows users to quantify the number, the volume and the intensity of the cells, in a number of different ways. For example, in Figure 3.4 **A**, an applet to run layer-wise analysis of the isocortex is shown. In Figure 3.4 **B**, an applet to create a cartoon-style heatmap showing the cell density or gene expression level is shown. Users can download the generated graphs in raster or vector format as well as the raw numerical values in CSV format. The notebook is intended for simple and quick inspection of the data, and the current collection of applets cover most of the common quantification tasks. If a custom analysis is needed, the user can simply download the reference-aligned brain data to the local machine and run customized analysis.

3.2.5 Sharing and publishing

The important concept of CUBIC-Cloud is sharing and publishing. With sharing, user can grant access to the brain data to other specific users. This feature is useful when exchanging data with internal or external research collaborators. User can choose read only or read and write access depending on how the data should be managed.

In addition to sharing with specific users, users can opt to publish their brain data in the CUBIC-Cloud's public repository. Once published, any users can view the brain. When a data is published, a hard copy of the data is created and registered in the database, to ensure the persistence of the data. Published data are given unique ID value so that the data can be accessed via URL or API request.

The share and publish capability are also supported for the notebooks and studios. Using this feature, users can transparently show their analysis results to the research community. To demonstrate this concept, over 60 brain data investigated in this study is deposited on CUBIC-Cloud public repository, as well as the notebooks and studios that performed the analysis.

3.2.6 Client APIs

The functionalities described so far all involves graphical interfaces, which would be useful to quickly and intuitively operate the software. CUBIC-Cloud also provides programmatic access to the service via REST APIs. These APIs would be useful for those who upload the brain data in a large batch, or those downloading brains from the public repository. In Table 3.1, some of the representative API endpoints are shown. Complete documentation of the CUBIC-Cloud API as well as the software development kit (SDK) for Python is under preparation as of this writing and will be made available soon.

3.3 Implementation of CUBIC-Cloud

CUBIC-Cloud's entire application stack is deployed on the cloud computing infrastructures offered by Amazon Web Service (AWS). The cloud is constructed using the

TABLE 3.1: Example list of CUBIC-Cloud API

API endpoint	Description	Parameters
`GET` /brains	Get a list of brains.	`type` (private, shared or public), `sortkey` (e.g. date or title), `keyword` (keyword search)
`POST` /brains	Upload a new brain.	NA
`GET` /brains/:item_id	Get a brain data by ID.	NA
`GET` /brains/downloads/ :item_id	Download selected raw data by brain ID.	`what` (raw or transformed cell table, raw or transformed structure image)
`GET` /brains/public/:item_id	Publish a brain by ID.	NA
`GET` /notebooks	Get a list of notebooks	`type` (private, shared or public), `sortkey` (e.g. date or title)
`POST` /notebooks/ :item_id/cells/cell_id	Run appelet task and generate a graph	(many parameters specifying the task)
`GET` /viewer/studios	Get a list of studio	`type` (private, shared or public), `sortkey` (e.g. date or title)
`GET` /viewer/studios/ launch/:item_id	Launch a studio	NA

serverless architecture (Adzic and Chatley, 2017). Serverless architecture has zero real (physical) server instances that are always running; instead, the cloud is composed by connecting microservices, which are dynamically invoked by events, and the computational resources are allocated by the cloud provider automatically on demand.

In designing CUBIC-Cloud, serverless approach was particularly advantageous compared to conventional cloud construction approaches in several ways. First, because CUBIC-Cloud is a scientific application, the concurrent access by the user would be not high (presumably up to tens of users at the same time), but the computational demand by each user could be potentially very high (for example, making large query to the database or submitting many brains to the cloud), meaning that the load on the system would be highly pulsatile. Therefore, the cloud system should be able to dynamically scale the computational power in response to a surge of the computational demand. In the conventional approach, such system would require a complex code base to dynamically launch or shutdown the physical servers, using frameworks like Kubernetes. By adopting serverless approach, this development cost is significantly mitigated, which is highly suitable for a small team of developers in scientific laboratories. Furthermore, serverless approach allows to flexibly divide the entire cloud system into a collection of microservices, in stead of a large monolithic application. Each microservices can independently request the CPU and memory resources as well as the runtime environment. Thus, the cloud system can be constructed in a modular manner, a welcoming feature for CUBIC-Cloud where a lot of prototyping is attempted.

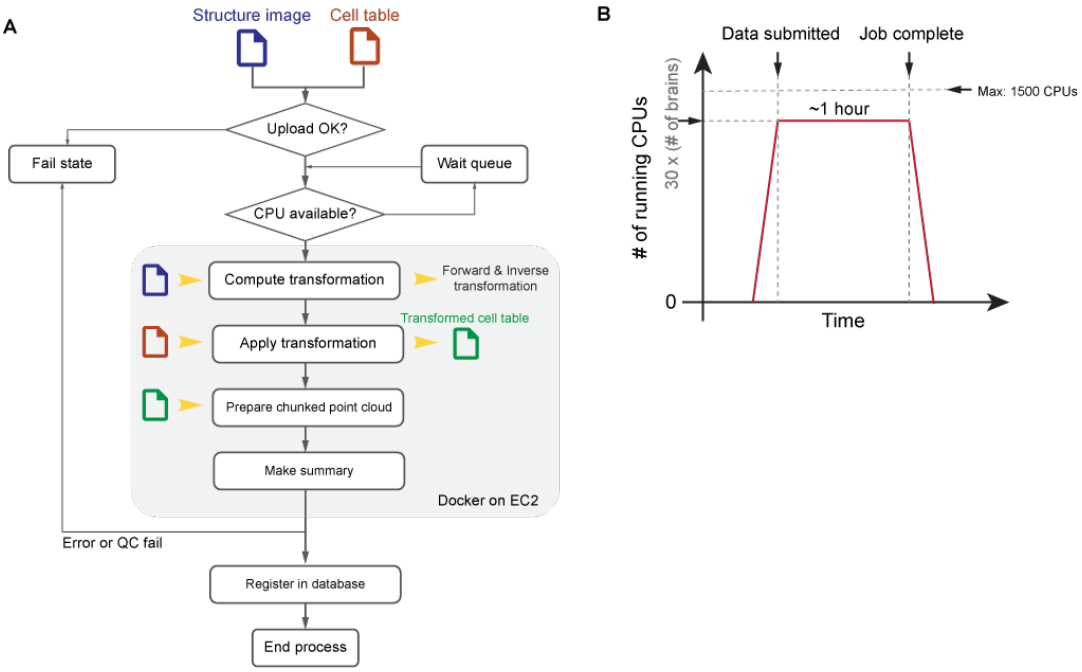

FIGURE 3.5: Implementation of the preprocessing task

A. Preprocessing task is orchestrated using AWS Step Functions framework. The schematic state transition model of the preprocessing is presented. **B**. The schematic diagram showing the cluster scaling.

The implementation details are described extensively in chapter 2.4.1. Here I particularly describe the automatic scaling of the preprocessing cluster, which is the core of the CUBIC-Cloud system. Preprocessing task is designed around the state machine model using AWS Step Functions framework (Figure 3.5 **A**). Once a user uploads the brain data, a upload completion event is fired from the storage system (S3). This event triggers the initiation of the state machine. First, Lambda (serverless compute unit provided in AWS) checks if the CPU allocation is below the limit imposed by AWS. Currently, standard AWS user account can request up to 1500 CPU cores at one time. If the running CPU core exceeds this limit, the task is placed in the wait queue, and job re-submission is attempted periodically. If the CPU allocation is available, a new compute instance (offered by AWS EC2) is launched. In the current implementation, *c5.9xlarge* instance type (36 CPU cores and 72 GB of RAM) is used. Although larger instance types are offered in EC2, ANTs registration program did not reduce execution time significantly if more than 40 CPU cores were used, presumably due to the memory access overheads. Thus, 36 CPU cores were determined to be optimal in terms of the execution time and the cost. As soon as the EC2 instance is launched and ready to use, a preprocessing program packaged in the Docker container is executed in the instance.

Preprocessing program first compute the registration between the submitted brain and the reference brain (see chapter 2.2.2), and outputs the transformation fields to map the coordinate from one brain to the other. Next, the computed transformation is applied to the cell table, so that the cell coordinates are mapped to the coordinate system of the CUBIC-Atlas. Then, a chunked point cloud data is generated for 3D visualization (see chapter 2.4.2). Lastly, a summary information of the analysis is created, and the brain is registered in the user's database.

The above preprocessing task takes about 1 hour of execution time. Once the preprocessing task is completed, the EC2 instance is quickly shutdown. When no preprocessing task is in the queue, the system CPU usage is zero, meaning that the idling cost is effectively zero. The schematic diagram showing the behavior of the cluster scaling is shown in Figure 3.5 **B**.

With this implementation, CUBIC-Cloud is able to process 41 (=1500/36) brains simultaneously, while consuming zero CPU during idle time. All other elements of the cloud system (such as fetching data or submitting notebook applet jobs) are also implemented with serverless construction (see chapter 2.4.1). Therefore, the entire cloud system can flexibly scale depending on the load.

3.4 Deployment and actual use cases

CUBIC-Cloud was initially released to the public in August 2020. Since then, several pilot users have signed up to try the cloud service. As an example use case, here I will describe a project where my colleague and I investigated the whole-brain parvalbumin (PV) expression (the result is reported in Chapter 4.2).

In this project, 60 whole-brain data were acquired (24 brains in the first batch and 36 brains in the second batch). After the data acquisition of the first batch (24 brains), these brains were uploaded to CUBIC-Cloud using a custom Python script utilizing CUBIC-Cloud's REST API. Upon the submission of 24 brains, the cloud system successfully scaled the cluster, resulting in 864 CPU cores running in parallel (24 $\times$ 36 CPUs). All data were processed within about 1.5 hours and the cluster successfully scaled out, leaving zero CPU core in the cluster. Without the cloud system, this calculation would have taken over a day, given that a single machine with 36 CPU cores is available as a local machine. Remarkably, thanks to the CUBIC-Cloud's infrastructure, such high-performance computing can be carried out without requiring much programming expertise. Indeed, in this project, my colleague (Mr. Kon), who does not have extensive programming skill, was able to analyze the whole-brain data with minimal effort. This use case demonstrates the usability and scalability of CUBIC-Cloud.

Chapter 4

Results 2: Analysis of Whole Mouse Brain Using CUBIC-Cloud

The construction of CUBIC-Cloud software was described in Chapter 3. CUBIC-Cloud is a novel software framework for analysis, visualization and sharing of whole mouse brain data. In this chapter, I will use CUBIC-Cloud framework to analyze various kinds of whole mouse brain data and demonstrate the utility of CUBIC-Cloud in neuroscience research. In chapter 4.1, I investigate the brain-wide distribution of five major neuronal and glial cell types visualized with whole-tissue immunostaining method. In chapter 4.2, I will expand the results of chapter 4.1 and explore the developmental changes of the PV-expressing neurons, ranging from juvenile to aged mouse brains. In chapter 4.3, the expression c-Fos gene is investigated to reveal the cellular clusters activated or repressed by pharmacological intervention. In chapter 4.4, I will study the brain-wide deposition of Aβ plaques of Alzheimer's disease model mouse. Lastly, the comprehensive mapping of the neural circuitry using rabies virus (RV) is demonstrated in chapter 4.5. These applications cover the major portion of the interest by the neuroanatomical studies, and thus show the generality of the CUBIC-Cloud framework.

Most of the data reported in this paper (except for the results described in chapter 4.2, which is unpublished) is deposited on CUBIC-Cloud public repository. The published data amounts to over 50 whole mouse brains, which provides a unique and comprehensive dataset for neuroscience research and demonstrates the concept of data sharing.

Note: The abbreviations of the brain region names are frequently used in this chapter, which follow the ontology defined by the Allen Brain Atlas. The look-up table can be found in Table 4.1.

4.1 Mapping whole-brain cell-type distribution

Contribution statement: The brain sampling, clearing and staining were done by M. Shimizu and K. Kon. LPS injection experiment described in chapter 4.1.4 was performed by K. Kon. LSFM imaging and data analysis were conducted by the author.

The gene expression (hence the cell-type distribution) in the brain is dynamically modulated by various developmental, behavioral and environmental factors, as discussed in the introduction (chapter 1.1.1). Understanding of the whole-brain state by means of the quantification of the gene expression at single cell resolution is thus of significant value. As the first application of CUBIC-Cloud, here I demonstrate the mapping of various cell-types in the adult mouse brain using 3D immunostaining

(Susaki et al., 2020). I studied the basal distribution of four neuronal subtypes of the wild-type mouse brain (chapter 4.1.1 to 4.1.2) and a subtype of glial cells, microglia (chapter 4.1.4). Together, these results present a comprehensive cellular map of the mouse brain, which would be a valuable resource for the neuroscientists. These brain datasets are openly available at the CUBIC-Cloud repository.

4.1.1 Whole-brain analysis of PV expressing cells and SST expressing cells

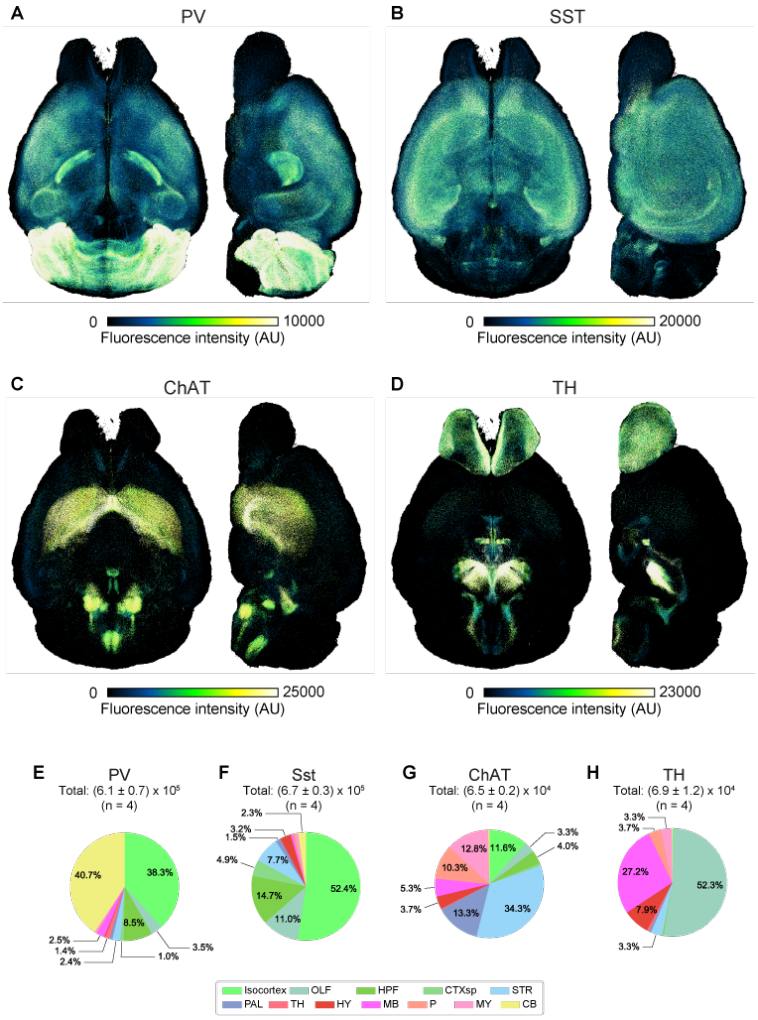

FIGURE 4.1: Whole-brain overview of the cell-type mapping

A-D. Whole-brain views of the investigated cell-types. The rendering was generated by CUBIC-Cloud's studio. Each dot (i.e. single cell) was assigned a pseudo-color based on its fluorescence intensity from the immunostaining image, reflecting the expression level. **E-H**. Pie chart showing the demography of the investigated cell-types in the major brain divisions. The ratio was computed by the number of the cells.

Parvalbumin-expressing (PV+) neurons and somatostatin-expressing (SST+) neurons are two major subtypes of the inhibitory neurons. Both PV+ and SST+ neurons underlie in essential cortical functions, including critical period regulation (Takesian and Hensch, 2013), learning (Donato, Rompani, and Caroni, 2013), and mental disorders such as schizophrenia (Lewis et al., 2012). Here, 8-weeks-old male C57BL/6N

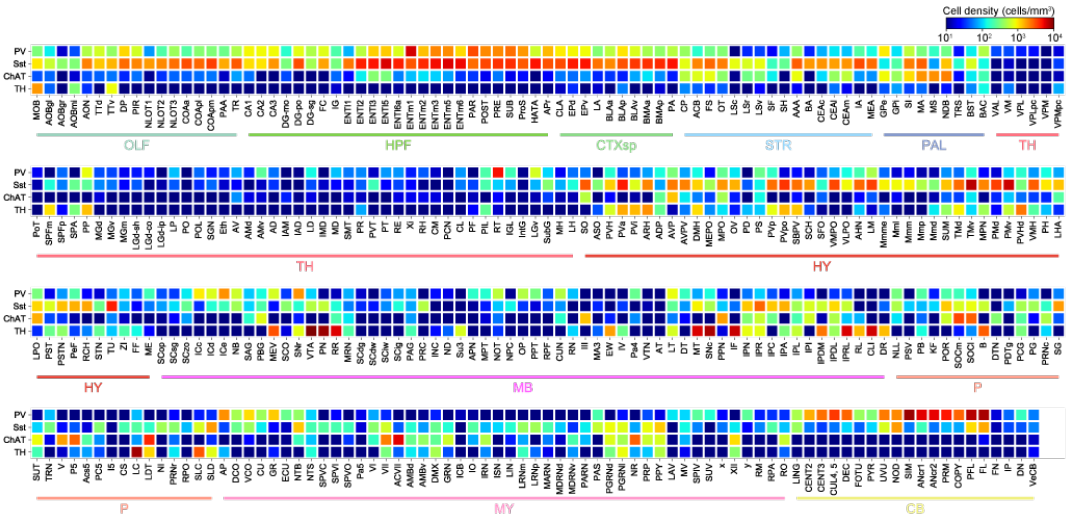

FIGURE 4.2: Density heatmap of the various cell-types across brain areas

The density of the investigated cell-types represented as a heatmap. Regions outside of the isocortex are shown here. Brain region acronyms follow the ontology defined by the Allen Brain Atlas.

mouse brain ($n = 4$) were cleared, imaged and analyzed following the methods described in Chapter 2. Brains were triple-stained with PV antibody conjugated with Cy3, SST antibody conjugated AlexaFluor 594 and nuclear staining dye (BOBO-1). LSFM images were obtained with (X,Y,Z) = (6.5, 6.5, 7.0) µm voxel resolution.

Figure 4.1 **A**, **B** presents the whole-brain overview of the PV+ and SST+ cell distribution, respectively. The total number of PV+ cells detected in my analysis was $(6.1 \pm 0.7) \times 10^5$, while SST+ cells amounted to $(6.7 \pm 0.3) \times 10^5$ (mean $\pm$ STD). While the largest majority of PV+ and SST+ neurons were found in the cortex, these cell types were universally distributed in the brain stem and in the cerebellum as well (Figure 4.1 **E**, **F**).

Within the isocortex, PV+ neurons were most densely populated in somatosensory and auditory areas ($\sim$ 3000-4000 cells/mm^3), followed by motor and visual areas ($\sim$ 2000 cells/mm^3) (Figure 4.3 **A, B**). PV+ neurons were scarce in the association areas (as low as 200 cells/mm^3), including ORB, PL and ILA. Compared to PV+ neurons, SST+ neurons were found across all isocortical areas with similar density ($\sim$ 3000-4000 cells/mm^3 (Figure 4.3 **D, E**). In terms of the cell density across layers, almost no PV cells were found in layer 1, and the PV+ cell density reached the maximum in layer 4 or 5 (Figure 4.3 **A,C**). SST+ neurons had a similar trend, and the density was highest in layer 4 or 5 (Figure 4.3 **D,F**).

In terms of the expression levels per cell, again, PV+ neurons showed large variance across cortical areas. As shown in Figure 4.4 **A** to **F**, some of the regions, like SSp-m and VISp, had a long tail in the distribution of the expression level, meaning that there is a population of neurons expressing a high amount of PV. On the other hand, regions like ILA and ECT had a very short tailed profile, meaning that the most of the cells have a weak PV expression. In a stark contrast, the expression distribution of SST+ neurons were quite homogeneous across cortical areas (Figure 4.4 **G** to **L**). This contrast between PV and SST rejects the possibility that the gene expression inhomogeneity of PV was an artifact from the staining or brain scanning. In the cortex, the PV expression level is known to be correlated with the glutamate

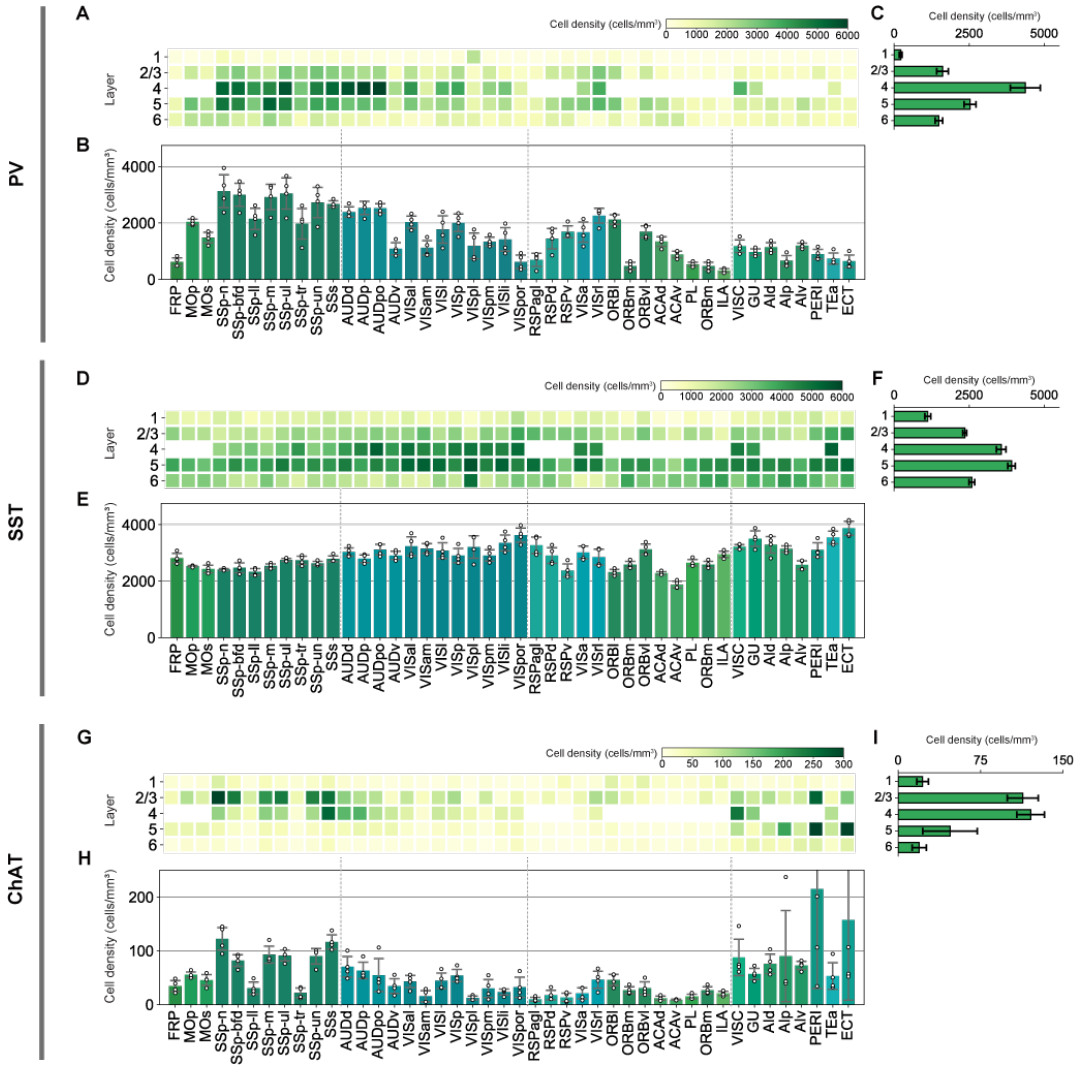

FIGURE 4.3: The density of the PV+, SST+ and ChAT+ neurons in the isocortex.

A,D,G. The density of the PV+, SST+, and ChAT+ neurons, respectively, in all areas in the isocortex is presented as a heatmap. **B,E,H**. The average across the layers of the data presented in **A**, **D**, **G**, respectively. **C,F,I**. The average across the regions of the data presented in **A**, **D**, **G**, respectively.

decarboxylase 67 (GAD67) expression level, which reflects the inhibitory power of the neuron (Donato, Rompani, and Caroni, 2013). Thus, the inhomogeneity of the PV expression may reflect the differences in the computation logic across different cortical areas.

The distribution of PV+ and SST+ cells in the subcortical areas are summarized as follows (Figure 4.2). Within the striatum, PV+ cells were observed with moderate density (a few hundred cells per mm^3), while they were almost entirely absent in the LS and the anterior, central, intercalated and medial amygdalar nucleus (AAA, CEA, IA and MEA). SST+ cells were universally observed in all areas in the striatum, with the average density of the striatum being 1100 cells/mm^3. The thalamus contained low numbers of PV+ cells, except that dense PV+ cells were present in the RT and PP. The thalamus contained low numbers of SST+ cells. In the hypothalamus, altough

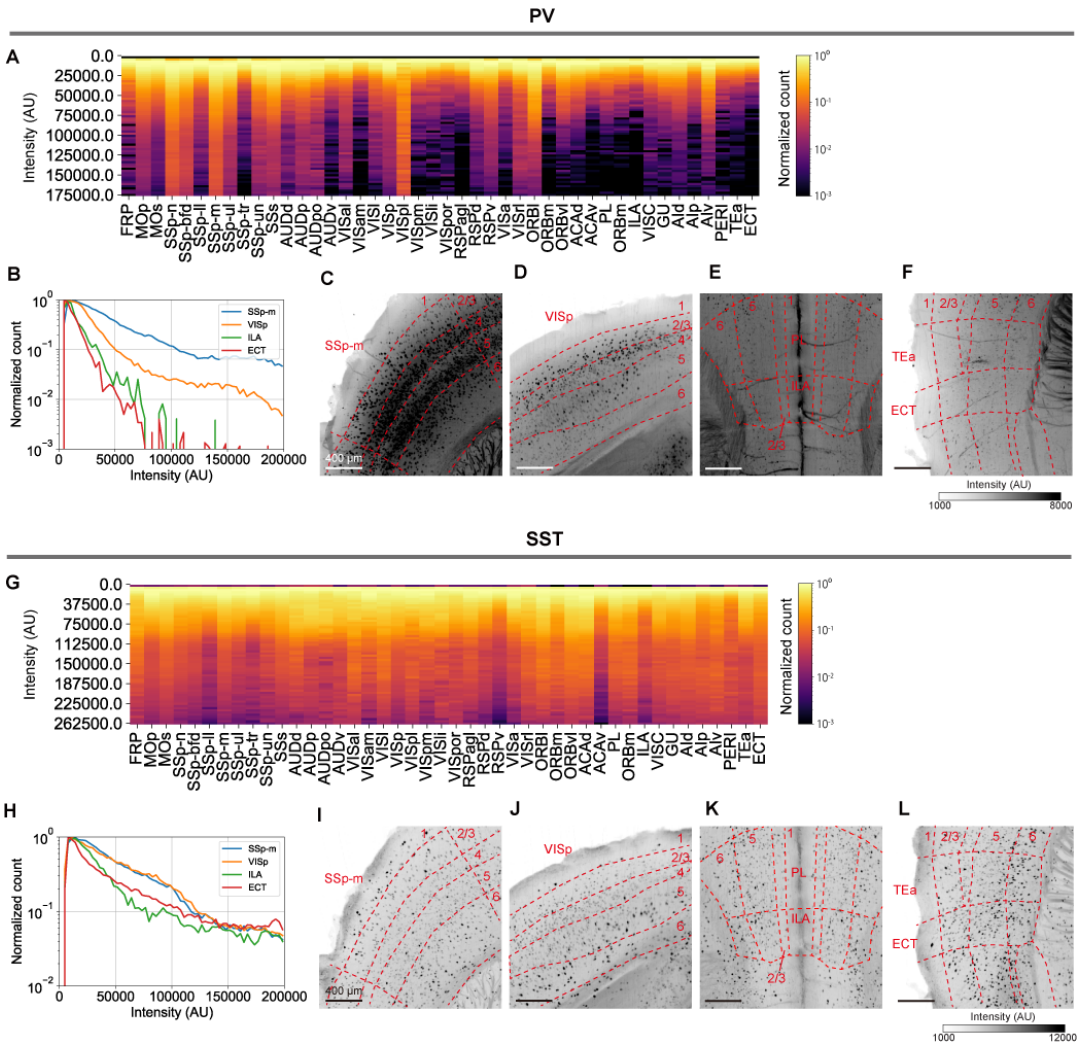

FIGURE 4.4: The expression level distributions in the major areas of the isocortex.

A. The distribution of PV expression levels per cell in the isocortex. Each column (i.e. the region) was normalized so that the peak equals to 1.0. **B**. The values of SSp-m, VISp, ILA and ECT were extracted from **A** and replotted as a distribution curve. **C-F**. Raw PV immunostaining images in SSp-m (**B**), VISp (**C**), ILA (**D**), and ECT (**D**). **G**. The distribution of SST expression levels per cell in the isocortex. Each column (i.e. the region) was normalized so that the peak equals to 1.0. **H**. The values of SSp-m, VISp, ILA and ECT were extracted from **A** and replotted as a distribution curve. **I-L**. Raw SST immunostaining images in SSp-m (**I**), VISp (**J**), ILA (**K**) and ECT (**L**).

PV+ cells were scarce, many nuclei contained medium to high density of SST+ neurons. Within the midbrain, PV+ cells were particularly abundant in the IC and SNr, while SST+ cells were most frequently observed in the RAmb. Within the pons and medulla, the NTB, SOC and NLL contained relatively high density of both PV+ and SST+ cells, while sparsely scattered populations were observed in other areas. In the cerebellum, there were a large number of PV+ neurons in Purkinje layers. Distinct SST+ cell clusters were found in the NOD and FL.

4.1.2 Whole-brain analysis of ChAT expressing cells

Next, I investigated the neurons expressing choline acetyltransferase (ChAT), a protein maker for the neurons that produce the neurotransmitter acetylcholine. Acetylcholine plays important signaling roles in the central nervous system (CNS), governing attention, learning and memory, sleep, and arousal (Everitt and Robbins, 1997). A degeneration of cholinergic neurons in the basal forebrain is recognized as a major hallmark of Alzheimer's disease (Mufson et al., 2008). 8-weeks-old male C57BL/6N mouse brain ($n = 4$) were cleared, imaged and analyzed following the methods described in Chapter 2. Brains were doubled-stained with ChAT antibody conjugated AlexaFluor 594 and nuclear staining dye (SYTOX-G). LSFM images were obtained with (X,Y,Z) = (6.5, 6.5, 7.0) µm voxel resolution.

The total number of ChAT+ cells detected in my analysis was $(6.5 \pm 0.2) \times 10^4$ (mean $\pm$ STD). Figure 4.1 **C** presents the whole-brain overview of the ChAT+ cell distribution. About half of ChAT+ cells were concentrated in the region collectively called the basal forebrain, which includes part of the striatum and pallidum (Figure 4.1 **G**). Continuously spreading from these regions, some ChAT+ cells were present in the hypothalamus, including lateral, medial and anteroventral preoptic areas (LPO, MPO, AVP) and SO. Other cholinergic neuron rich regions included PPN of the midbrain and LDT of the pons. In addition, ChAT+ neurons were aggregated in cranial nerve nucleus, including oculomotor nucleus (III), motor nucleus of trigeminal (V), abducens nucleus (VI), facial motor nucleus (VII) and hypoglossal nucleus (XII) (Figure 4.2). Within the isocortex, there were sparse (less than 100 cells/mm^3) populations of ChAT+ neurons. These neurons expressed very low amount of ChAT, compared to ChAT+ cells in the subcortical regions. These ChAT+ neurons were most dense in layer 2/3 or 4 (Figure 4.3 **G** and **I**).

4.1.3 Whole-brain analysis of TH expressing cells

Next, I investigated the neurons expressing tyrosine hydroxylase (TH), a protein maker for the catecholamine-secreting neurons, which include neurotransmitters dopamine, adrenaline and noradrenaline. 8-weeks-old male C57BL/6N mouse brain ($n = 4$) were cleared, imaged and analyzed following the methods described in Chapter 2. Brains were doubled-stained with TH antibody conjugated AlexaFluor 594 and nuclear staining dye (SYTOX-G). LSFM images were obtained with (X,Y,Z) = (6.5, 6.5, 7.0) µm voxel resolution.

The total number of Th+ cells detected in my analysis was $(6.9 \pm 1.2) \times 10^4$ (mean $\pm$ STD). Figure 4.1 **D** presents the whole-brain overview of the TH+ cell distribution. The majority of the detected TH neurons were localized in well-known dopaminergic cell groups (A8 to A16) and noradrenergic cell groups (A1 to A7) (Dahlström and Fuxe, 1964). Dopaminergic cell groups include the RR, SNc, rostral and central linear nucleus raphe (RL and CLI) and VTA, which form the A8, A9 and A10 in the midbrain. Within the hypothalamus, TH neurons were clustered in the periventricular hypothalamic nucleus, anterior, posterior, intermediate and propotic parts (PVa, PVp, PVi, PVpo), ARH, ZI and ADP, which form A11-A15 cell groups. TH neurons were numerous in the olfactory area (A16), selectively localized in the glomerular layer. Noradrenergic cell groups formed distinct bands crossing several nuclei in the medulla and pons, which included the LRN, NTS and DMX, which form A1 and A2. In the pons, a particularly high density was observed in and around LC, which forms A6. No significant population of TH+ cells were observed in the isocortex, hippocampus, cortical subplate, striatum and pallidum.

4.1.4 Whole-brain analysis of Iba1 expressing cells

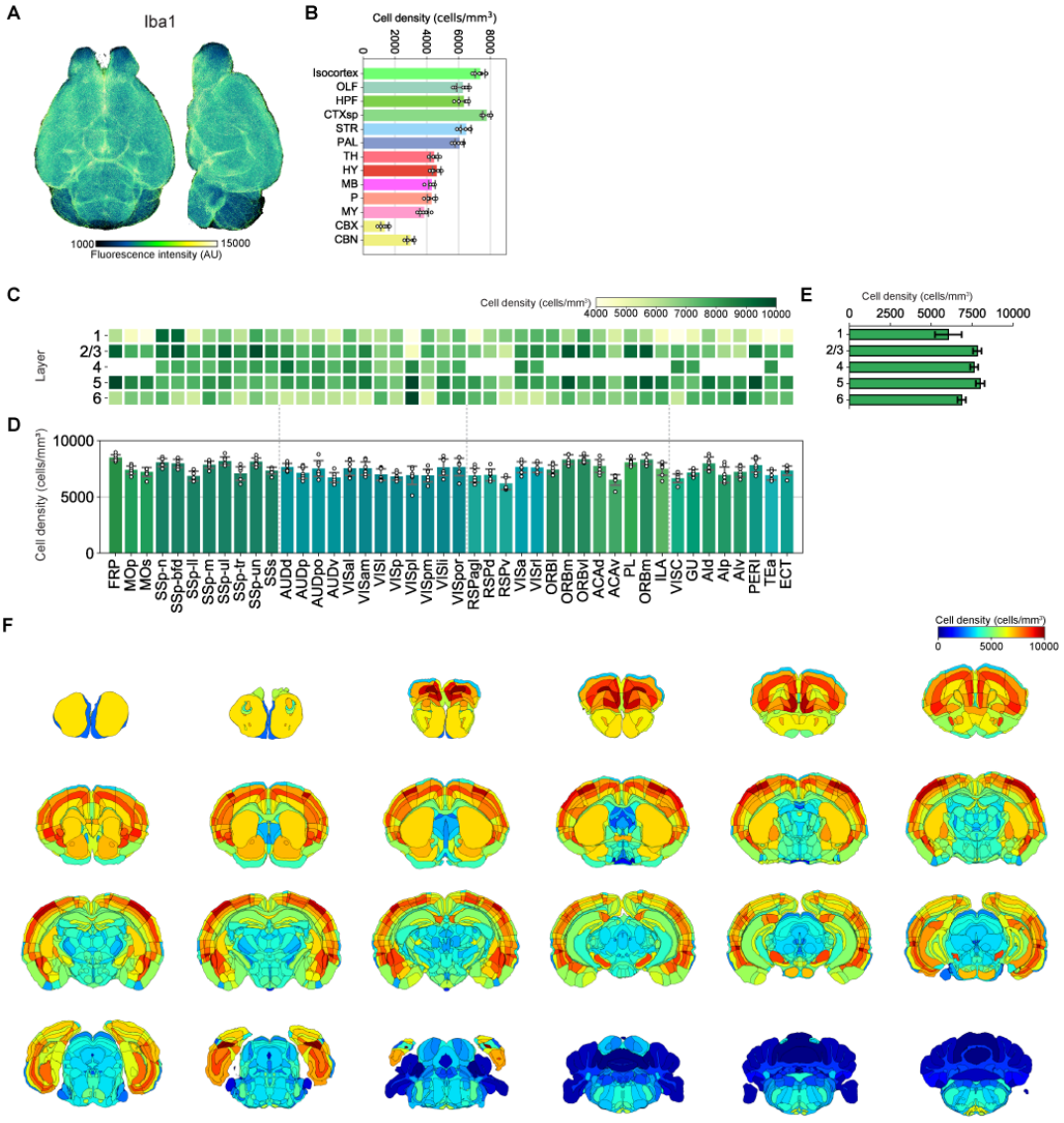

FIGURE 4.5: The distribution of Iba1+ in the whole mouse brain.

A. Whole-brain view of all Iba1+ cells. Each dot (i.e. single cell) was assigned a pseudo-color based on its fluorescence intensity from the immunostaining image, reflecting the expression level. **B**. Density of Iba+ cells in major brain divisions. **C**. The density of the Iba1+ cells in the isocortex is presented as a heatmap. **D**. The average across the layers of the data presented in **C**. **E**. The average across the regions of the data presented in **C**. **F**. Cartoon-style heatmap showing the Iba1+ cell density across all brain regions.

Ionized calcium-binding adapter molecule 1 (Iba1) is a cell-type marker for microglia. Microglia plays an key role in the active immune defense in the CNS, among many other important functions. As a model system to study the immune response of the body and the brain, lipopolysaccharides (LPS), a purified extract of the outer membrane of Gram-negative bacteria, is often used. In this chapter, I will describe an experiment where I investigated the gene expression change of microglia upon LPS-induced inflammation.

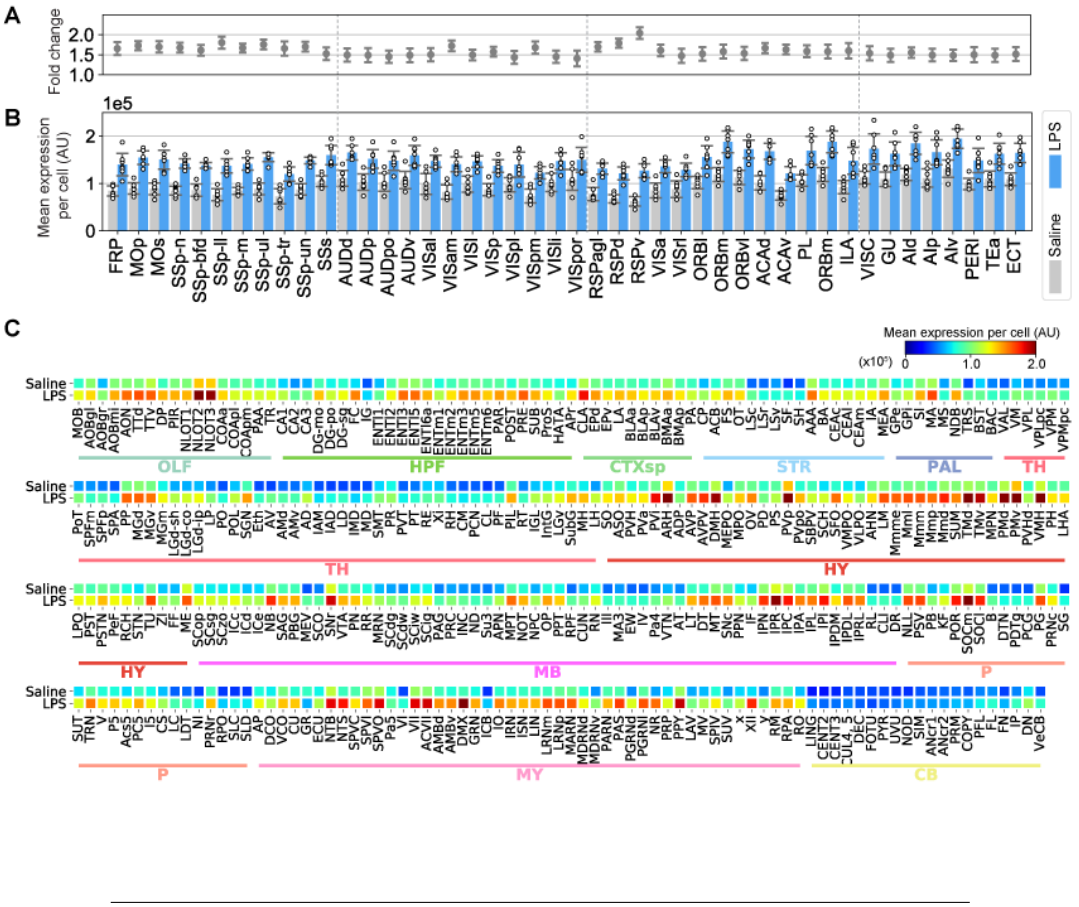

FIGURE 4.6: Mean Iba1 expression level per cell, comparing saline- and LPS-administered conditions.

A,B. The mean Iba1 expression level per cell in the isocortex, comparing saline- and LPS-administered conditions. Fold change in the expression level ($I_{\mathrm{LPS}}/I_{\mathrm{saline}}$) is shown in **A**. **C**. Heatmap showing mean Iba1 expression level per cell in all brain regions outside the isocortex, comparing saline- and LPS-administered conditions.

In the experiment, 1 mg/kg of LPS was administered to mice via intraperitoneal (i.p.) injection ($n = 7$). The control group was administered saline ($n = 7$). Twenty-four hours after injection, the brains were dissected. Then, the brains were cleared, imaged and analyzed following the methods described in Chapter 2. Brains were doubled-stained with Iba1 antibody and nuclear staining dye (SYTOX-G). LSFM images were obtained with (X,Y,Z) = (6.5, 6.5, 7.0) µm voxel resolution.

I first describe the observation on normal (i.e. saline-administered) Iba1 expressing (Iba1+) cell distribution. The total number of Iba1+ cells detected in my analysis was $(2.72 \pm 0.14) \times 10^6$ ($n = 7$), ubiquitously residing in all brain areas (Figure 4.5 **A**). The average cell density was highest in the cortical areas (6000 to 8000 cells/mm^3), intermediately dense in the brain stem ($\sim$ 4500 cells/mm^3) and lowest in the cerebellum (Figure 4.5 **B**). Overall, the standard deviations of the Iba1+ cell density between animals were usually as small as 5%, implying the highly regulated microglial proliferation. In the isocortex, all areas had almost same density of Iba1+ cells (Figure 4.5 **C,D**). In terms of the laminar structures, layer 1 and layer 6 had slightly lower density, while layer 2/3, 4 and 5 contained almost same density (Figure 4.5 **C,E**). In the subcortical areas, some regions had particularly high Iba1+ cell density, including amygdala nucleus, globus pallidus, external segment (GPe) and substantia nigra,

reticular part (SNr) (Figure 4.5 **F**).

I next compared Iba1 expression levels, as well as absolute Iba1+ cell count, between LPS- and saline-administered groups. In the isocortex, no areas displayed significant change in the Iba1+ cell density ($p > 0.1$), however, the mean expression level per cell was increased significantly in all areas ($p < 0.05$) (Figure 4.6 **B**). The elevated Iba1 expression is known to be correlated with microglial activation (Ito et al., 1998). Intriguingly, the fold change in terms of the expression level was almost constant across all areas, sitting at around 1.5 (Figure 4.6 **A**). One region, RSPv, had slightly higher increase (2.0-fold) than other areas. This may be due to the fact that RSPv is neighboring to the third ventricle (v3), where cytokines produced from the inflammation may be more concentrated in cerebrospinal fluid (CSF).

Outside the isocortex, most regions showed significant increases in Iba1 expression amount (Figure 4.6 **C**), revealing the brain-wide response to the inflammation state. It was confirmed that the increase in Iba1 expression was markedly high in the SFO and IO, which are part of ircumventricular organs (CVOs) having highly permeable blood-brain barrier (BBB) (Furube et al., 2018). On the other hand, Iba1 expression amount barely changed in cerebellum.

4.2 Mapping whole-brain neural development: The PV neurons

Contribution statement: The brain sample preparation, clearing and staining was done by K. Kon and the author. LSFM imaging and data analysis was conducted by the author.

In chapter 4.1, the distribution of major neuronal and glial cell types were investigated using whole-brain imaging and CUBIC-Cloud analysis framework. The results presented were all collected from adult ($\sim$ 8-weeks-old) mouse brains. Crucially, the landscape of the gene expression within the brain dynamically changes throughout the course of development and aging, which calls for the developmental analysis of the gene expression. However, the amount of data required to trace the developmental changes increases linearly with the number of time points, making the brain-wide survey of the developmental gene expression an extremely difficult task. The high-throughput imaging enabled by LSFM and scalable data analysis powered by cloud computing may provide an effective solution to this challenge. A distributed and collaborative data acquisition facilitated by cloud-based data sharing would further accelerate the research. To demonstrate such possibility, here I will investigate the whole-brain development of the parvalbumin (PV) expressing neurons.

To survey a comprehensive expression pattern of PV neurons covering the mouse's entire life span, the C57BL/6N wild-type mouse brains from 10 different ages were collected[1](3-, 4-, 5-, 6-, 8-, 10-, 12- and 14-weeks, 8- and 24-months-old males). Following the methods described in chapter 2, the brains were cleared and triple-stained with PV antibody conjugated with AlexaFluor 594, Iba1 antibody and nuclear staining dye (SYTOX-G). Here, only the results of PV neurons are presented, and the results of Iba1 cell analysis will be published elsewhere. LSFM images were obtained with (X,Y,Z) = (6.5, 6.5, 6.5) µm voxel resolution.

[1]The analysis of 1-weeks-old and 2-weeks-old brains were attempted but was not successful, due to the difficulty of the immunostaining and the fact that the PV neurons in the cortex are born at the end of the second postnatal week (Bitzenhofer, Pöpplau, and Hanganu-Opatz, 2020).

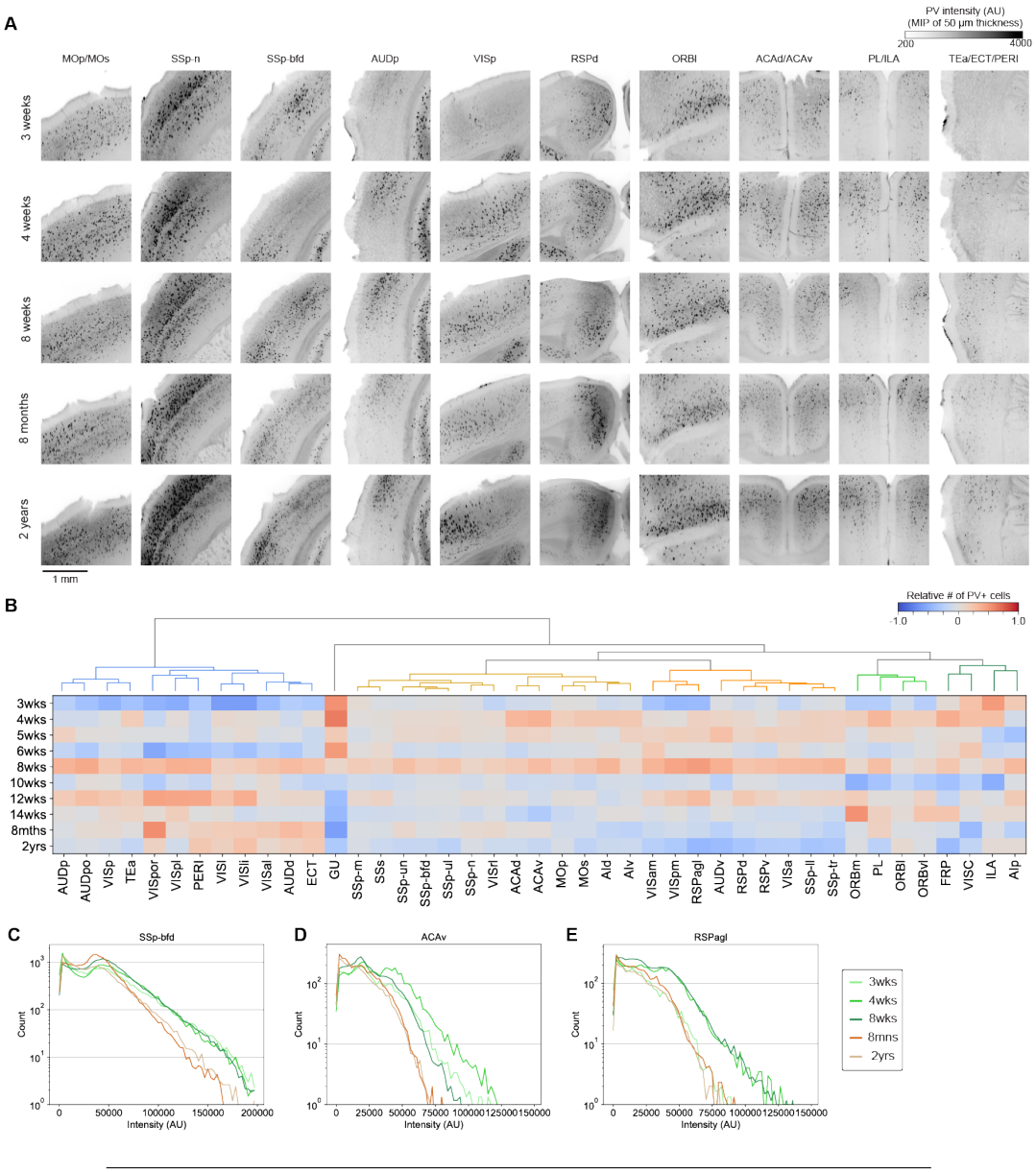

FIGURE 4.7: The PV expression within the cortical areas across the mouse brain development.

A. The representative PV immunostaining images showing ten cortical areas at different developmental stages. **B**. Relative number of PV+ cells in the isocrotex. For the raw cell count $C_{i,j}$ of region j at age i, the relative cell count $\tilde{C}_{i,j}$ was calculated as $\tilde{C}_{i,j} = (C_{i,j} - \frac{1}{N}\sum_k C_{k,j})/\frac{1}{N}\sum_k C_{k,j}$. Then, the columns (i.e. the brain regions) were clustered using Ward's method. **C, D, E**. The distribution of PV immunostaining signal intensity per cell, computed over three cortical areas, SSp-bfd (**D**), ACAv (**D**) and RSPagl (**E**), respectively. For the color and age correspondence, refer to the inset of the panel.

Figure 4.7 **A** shows the raw PV immunostaining images of selected cortical areas at different ages. As shown in these images, the PV expression in general is kept at similar levels across different ages, spanning from juvenile (3-weeks-old) to aged (2-years-old). This observation is quantified in Figure 4.7 **B**, where the relative number of PV+ cells are presented. The columns (i.e. the brain regions) were hierarchically clustered using Ward's method, which revealed several distinct categories. In the

first cluster (blue), which included visual (VIS) and auditory (AUD) cortex, I observed a trend that the number of PV+ cells are relatively low in 3-weeks-old brain, after which the number stays at almost the same level. The second large cluster (yellow and orange) contained the motor (MO), sensory (SS) and retrosplenial (RSP) areas, where the number of PV+ cells stayed almost constant from 3-weeks-old to 2-years-old. The orange and yellow clusters differed in that the slight decrease at 2-years-old brain was observed in the orange cluster. The third cluster (light and dark green) was rather noisy, but there is a small transient increase in the number of PV+ cells at 4-weeks-old. In addition to the number of PV+ cells, I also investigated the PV expression amount per cell in each brain area, revealing some areas showing drastic changes. For example, in the SSp-bfd, the PV expression decreased in 8-months- and 2-years-old brain (Figure 4.7 **C**). In the ACAv and RSPagl, the PV expression transiently increased in 4-weeks-old, then decreased as aging (Figure 4.7 **D**).

Compared to the cortical areas, the PV gene expression in the subcortical areas changed more drastically across the mouse life span. The reticular nucleus of the thalamus (RT) is populated with a dense PV+ neurons. The analysis revealed that the number of PV+ cells in RT gradually decreased from 3-weeks-old until the 2-years-old (Figure 4.8 **C**, **D**). In the zona incerta (ZI), the number of PV+ cells rapidly decreased from 3-weeks-old to 5-weeks-old, and stabilized afterwards (Figure 4.8 **E**, **F**). In the anterior pretectal nucleus (APN), the number of PV+ cells continuously decreased until 2-years-old, at which point PV+ cells were almost entirely absent (Figure 4.8 **G**, **H**). The globus pallidus, external segment (GPe) is another major hub of PV+ neurons, and the number of PV+ neurons in GPe slightly decreased over aging, but not as significantly as RT, ZI or APN (Figure 4.8 **A**, **B**). Lastly, the decrease in the PV+ cells were also observed in the inferior colliculus (IC), which is a relay nuclei of auditory inputs (Figure 4.8 **I**, **J**). The decrease was more evident in the ventral portion of the IC, while the superficial layer PV+ neurons were not strongly affected.

Intriguingly, RT, ZI and APN are functionally related in that these nuclei send inhibitory inputs to the first-order and higher-order thalamic nuclei (Giber et al., 2008; Li et al., 2020), while GPe primarily projects to non-thalamic nuclei including STN and SNr (Park et al., 2019). Thus, the decrease in PV+ neurons over development in RT, ZI and APN may reflect the changes in the sensory processing and gating.

The RT is a relatively large nuclei having spatially heterogeneous functions and projection patterns. A close inspection revealed that the changes in PV+ expression in the RT was spatially heterogeneous (Figure 4.9 **A**). The decrease in the PV+ cell number as well as the PV expression levels were most pronounced at around the middle part of the RT along anterior-posterior axis (Figure 4.9 **B**, **C**, **D**, **E**). On the other hand, the anterior and posterior end of the RT, respectively, showed relatively small decrease in PV. According to a recent study (Li et al., 2020), there are distinct two types of neurons in RT, which selectively project to first-order and higher-order thalamic nuclei, respectively. These neurons can be roughly differentiated by the expression of *Spp1* and *Ecel1* proteins, respectively. In the future studies, it would be interesting to investigate how much of these neuron types are lost during development and aging.

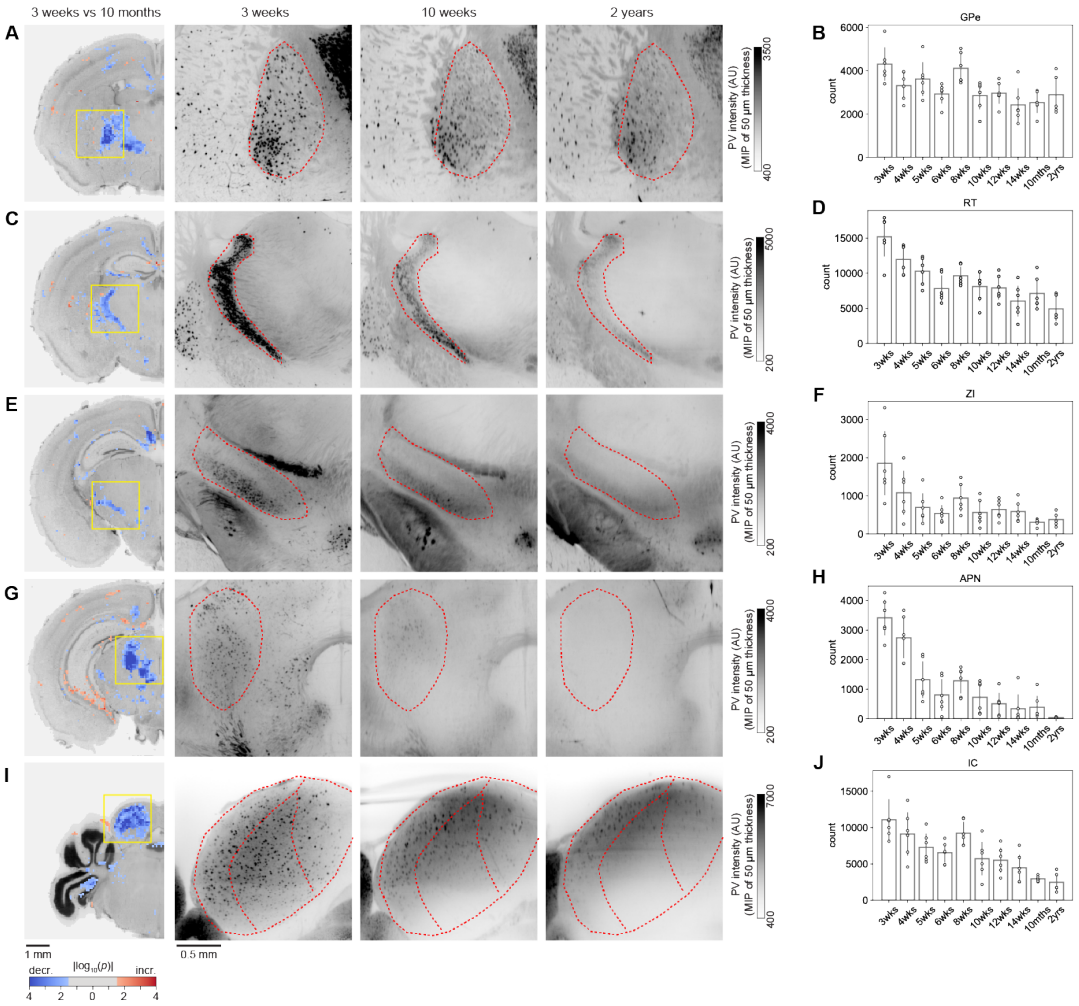

FIGURE 4.8: The PV expression within the subcortical areas across the mouse brain development.

The panels of the leftmost column shows the p value map calculated by comparing the number of PV+ cells of 3-weeks-old and 10-months-old mouse brains. The following three columns show the raw PV immunostaining images corresponding to the boxed regions in the p value map, taken from 3-weeks-old, 10-weeks-old and 2-years-old mouse brain, respectively. The rightmost panels show the number of PV+ cells in the corresponding brain regions. **A, B**. Globus pallidus (GPe). **C, D**. Reticular nucleus of the thalamus (RT). **E, F**. Zona incerta (ZI). **G, H** Anterior pretectal nucleus (APN). **I, J** Inferior colliculus (IC).

4.3 Mapping whole-brain neuronal activity profile using IEGs labeling

Contribution statement: LPS injection was performed by K. Kon and the author. The brain clearing and staining was done by K. Kon. LSFM imaging and data analysis were conducted by the author.

The next important application domain of CUBIC-Cloud is to reconstruct the neuronal activity profile by imaging the protein expressions of immediate early genes (IEGs) such as c-Fos. As was described in the introduction (Chapter 1.2.3), such automated analysis would allow comprehensive identification of cellular clusters that

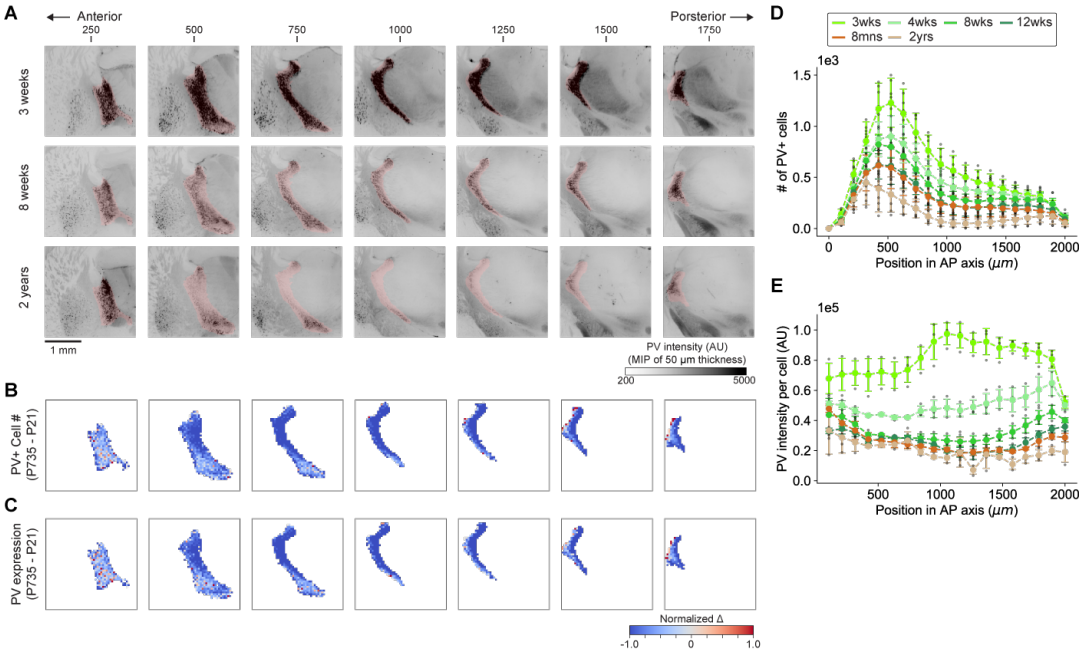

FIGURE 4.9: The PV expression within the RT across the mouse brain development.

A. Raw PV immunostaining images showing the RT at consecutive coronal positions. The regions highlighted by red is the RT. **B**. The heatmap showing the difference of the number of PV+ cells between 2-years-old and 3-weeks-old mouse brain. The difference was normalized as $\Delta = (C_{2\text{yrs}} - C_{3\text{wks}})/C_{3\text{wks}}$. **C**. The heatmap showing the difference of the PV expression level between 2-years-old and 3-weeks-old mouse brain. The difference was normalized in the same way as in **B**. **D**. The number of the PV+ cells within the RT at each coronal positions. **E**. The mean PV expression level per cell within the RT at each coronal sections. For the color and age correspondence, refer to the inset of the panel.

underlie an animal's behavioral phenotype. Reciprocally, one could define an animal's phenotype in a bottom up manner based on the activity pattern of neuron ensembles.

Particularly, IEG-based activity reconstruction would be suitable to track relatively slow neural dynamics, such as wake-sleep cycles. As a model system to explore this application, here I studied the IEG profile under the administration of lipopolysaccharides (LPS). As described in Chapter 4.1.4, LPS administration induces the acute inflammation, accompanied with a prolonged sleep which lasts for several hours. The neural mechanism causing the sleep behavior upon LPS-induced inflammation is not fully elucidated yet. Thus, I investigated the whole-brain IEG profile under LPS administration.

150 µg/kg LPS were administered to mice via intraperitoneal injection at CT = 14, and the brains were sampled between CT = 16 and 17. Following the methods described in Chapter 2, the brains were cleared and double-stained with c-Fos antibody conjugated with AlexaFluor 594 and nuclear staining dye (SYTOX-G).

Figure 4.10 **A** shows the whole-brain 3D rendering of the detected c-Fos expressing cells. I comprehensively searched for the activated or repressed brain regions by both region-wise and voxel-wise statistical analysis. As a result, it was found that the c-Fos expressions in some of the isocortical areas were reduced, which included motor and somatosensory areas, presumably reflecting the mouse's resting

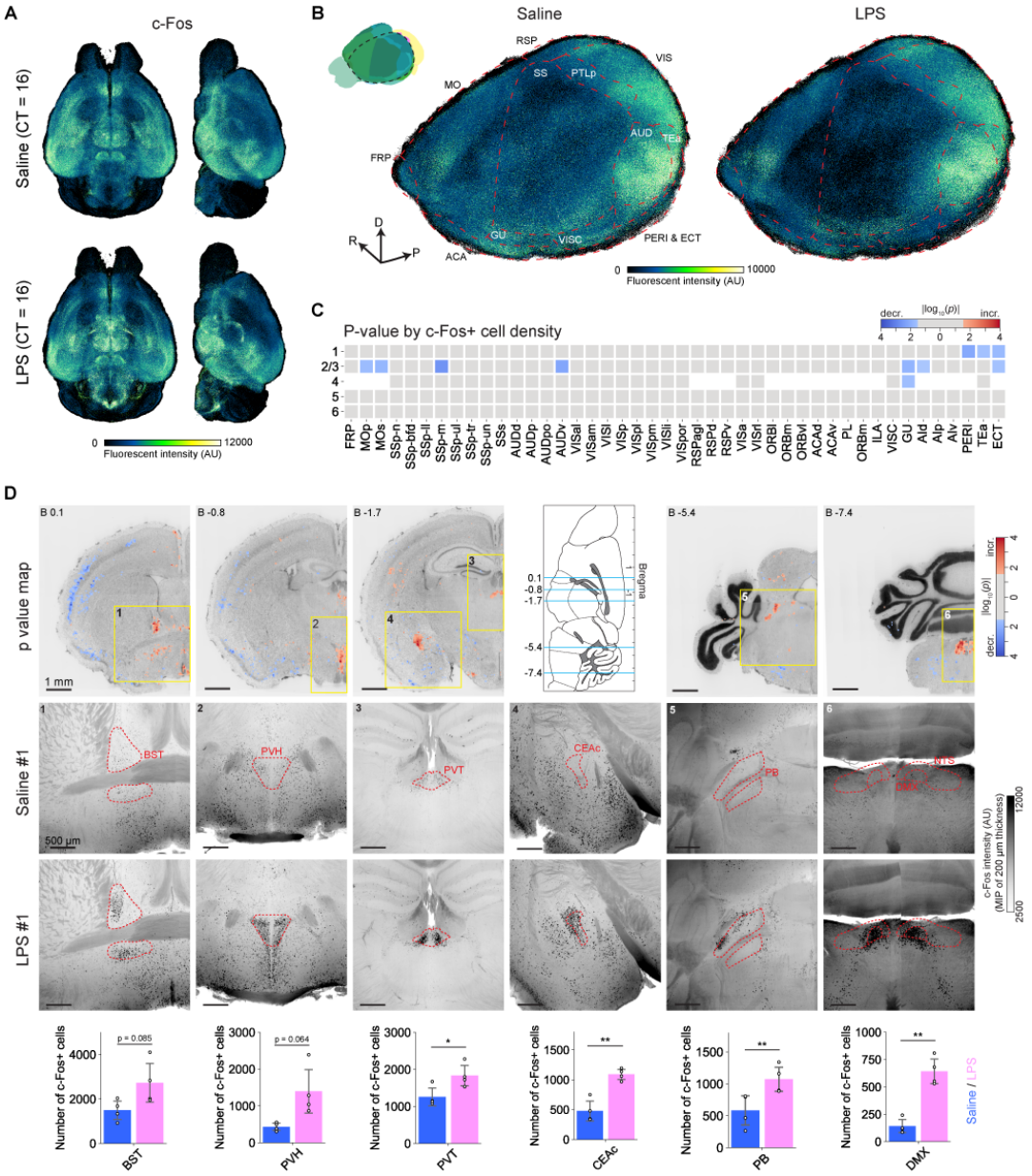

FIGURE 4.10: Whole-brain analysis of c-Fos expression level changes induced by LPS administration

A. Whole-brain views of c-Fos+ cells, showing saline (upper) and LPS (lower) administered brains. Each point (i.e. single cell) was assigned a pseudo-color based on its fluorescence intensity. **B**. Magnified 3D view of **A**, where the left isocortex was selectively displayed. Orientation arrows stand for R (right), D (dorsal) and P (posterior). **C**. P-value heatmap showing the isocortex regions whose c-Fos+ cell density was significantly affected by LPS. P-value was computed by comparing the c-Fos+ cell count. The color lookup table is log scaled (base 10), where red color represents the regions that were activated (i.e. more c-Fos+ cells) by LPS, and blue represents the repressed regions. Regions with no statistical significance ($p > 0.05$) were assigned a gray color. **D**. Distinct brain regions activated by LPS. The top row shows the voxel-wise p-value map. Color lookup table follows that of **C**. The second and third rows are the raw c-Fos images of saline- and LPS-administered group, respectively. The forth row shows the number of c-Fos+ cells of the identified regions. *$p < 0.05$, **$p < 0.01$, Welch's t-test. See Table 4.1 for the definitions of brain region acronyms.

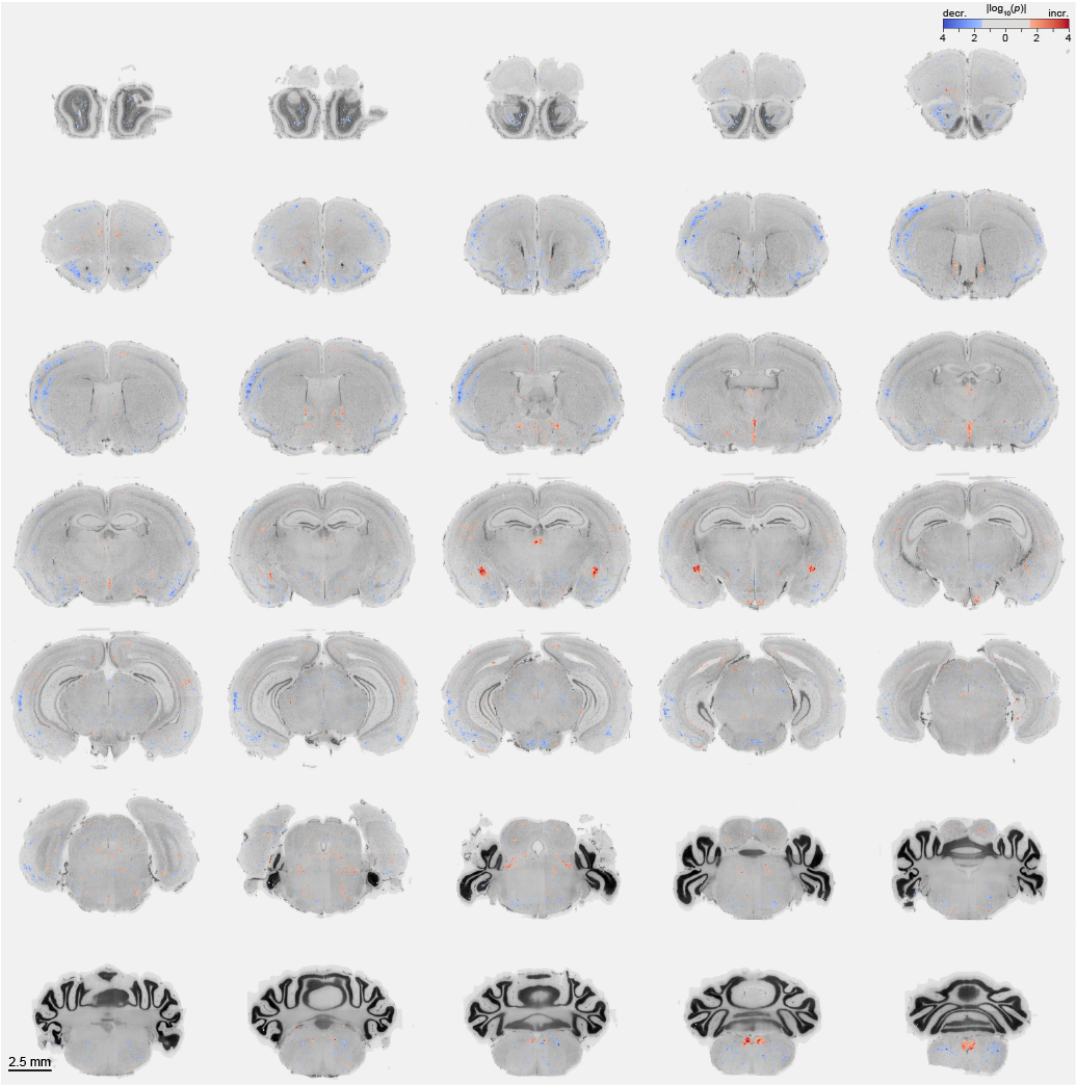

FIGURE 4.11: Voxel-wise p-value heatmap showing the affected regions by LPS.

p-values of individual voxels were computed by c-Fos+ cell count between saline- and LPS-administered groups. The color lookup table is log scaled (base 10), where red color represents the regions that were activated (i.e. more c-Fos+ cells) by LPS, and blue represents the repressed regions. Voxels with no significance ($p > 0.05$) were uncolored. Background is nuclear staining image of CUBIC-Atlas for navigation. Step between consecutive slices is 0.34 mm.

state (Figure 4.10 **B** and **C**). I also found that some distinct brain nuclei were activated by LPS. Among those, the most notable regions included the bed nuclei of the stria terminalis (BST), paraventricular hypothalamic nucleus (PVH), paraventricular nucleus of the thalamus (PVT), central amygdalar nucleus (CEA), parabrachial nucleus (PB), nucleus of the solitary tract (NTS) and dorsal motor nucleus of the vagus nerve (DMX) (Figure 4.10 **D**). BST, PVH and CEA share common functions in that they respond to stress exposure through intricate interactions (Hsu et al., 1998; Choi et al., 2007). NTS and DMX receive inputs from the vagal nerves, which would transmit the inflammation-induced signals from the gut nerve.

Within the BST, a specific subdivision (the oval region; ovBST) was strongly activated by LPS (Figure 4.10 **D**). Indeed, according to a recent study, ovBST is responsible for the inflammation-induced aronexia (Wang et al., 2019). ovBST receives inputs from CEA and PB, which were also found to be activated in my analysis. Therefore, the present result was able to successfully identify elevated c-Fos expressions in these spatially separated yet functionally related regions.

I also observed heterogeneous c-Fos activation in the PVT. In terms of the number of c-Fos+ cells, the increase in number was more pronounced in the posterior PVT (pPVT) than the anterior PVT (aPVT) (Figure 4.12, **A**, **B** and **C**). In terms of the expression level, both pPVT and aPVT showed similar level of increase (Figure 4.12 **D** and **E**). Recently, Gao et al. (Gao et al., 2020) identified two classes of distinct neurons in PVT. Type I neurons, densely located in the pPVT, responds to aversive stimuli. On the other hand, type II neurons, dominantly located in the aPVT, become silent upon aversive stimuli. It is also reported that the type II neurons were active during sleep. In the pPVT, our observation aligns with the insight by Gao, where the activated population was likely Type I neurons. In the aPVT, our result might reflect the mixed response of the type II neurons, where aversive inflammatory stimuli and induced sleep were both present. It should also be noted that Type I neurons in the pPVT project to CEA, ILA and ACB. I indeed observed that ILA and ACB were weakly activated (Figure 4.11).

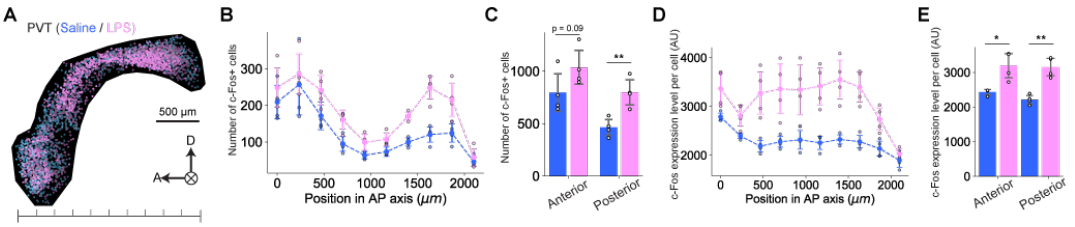

FIGURE 4.12: Whole-brain Analysis of c-Fos Expression Level Changes by LPS Administration

A. Plot of c-Fos+ cells in PVT. Cells are pseudo-colored with their intensity values. Pink (blue) dots are from LPS (saline) administered brains, respectively. **B**. The number of c-Fos+ cells in the PVT in 10 divisions along the anterior-posterior (AP) axis. **C**. The number of c-Fos+ cells in the anterior and posterior half of the PVT. The boundary between anterior and posterior region was set at the center of the PVT along AP axis. **D**. The c-Fos expression levels per cell in the PVT in 10 divisions along the anterior-posterior (AP) axis. **E**. The c-Fos expression levels per cell in the anterior and posterior half of the PVT. *$p < 0.05$, **$p < 0.01$, Welch's t-test.

4.4 Whole-brain analysis of Alzheimer's disease model mouse

Contribution statement: The brain preparation, clearing and staining were done by H. Ono. LSFM imaging and data analysis were conducted by the author.

I next applied CUBIC-Cloud analysis framework to quantitatively understand the pathological state of the Alzheimer's disease (AD) model mouse. To demonstrate this, the whole brain from an *App*$^{\text{NL-G-F/NL-G-F}}$ AD model mouse (Saito et al., 2014) (9- to 10-months-old) was cleared and stained by anti-Aβ antibody ($n = 4$). In LSFM images, Aβ plaques were observed as dim blobs often accompanying a bright spot at the core. Although the plaques are not cells, their blob-looking appearance allowed

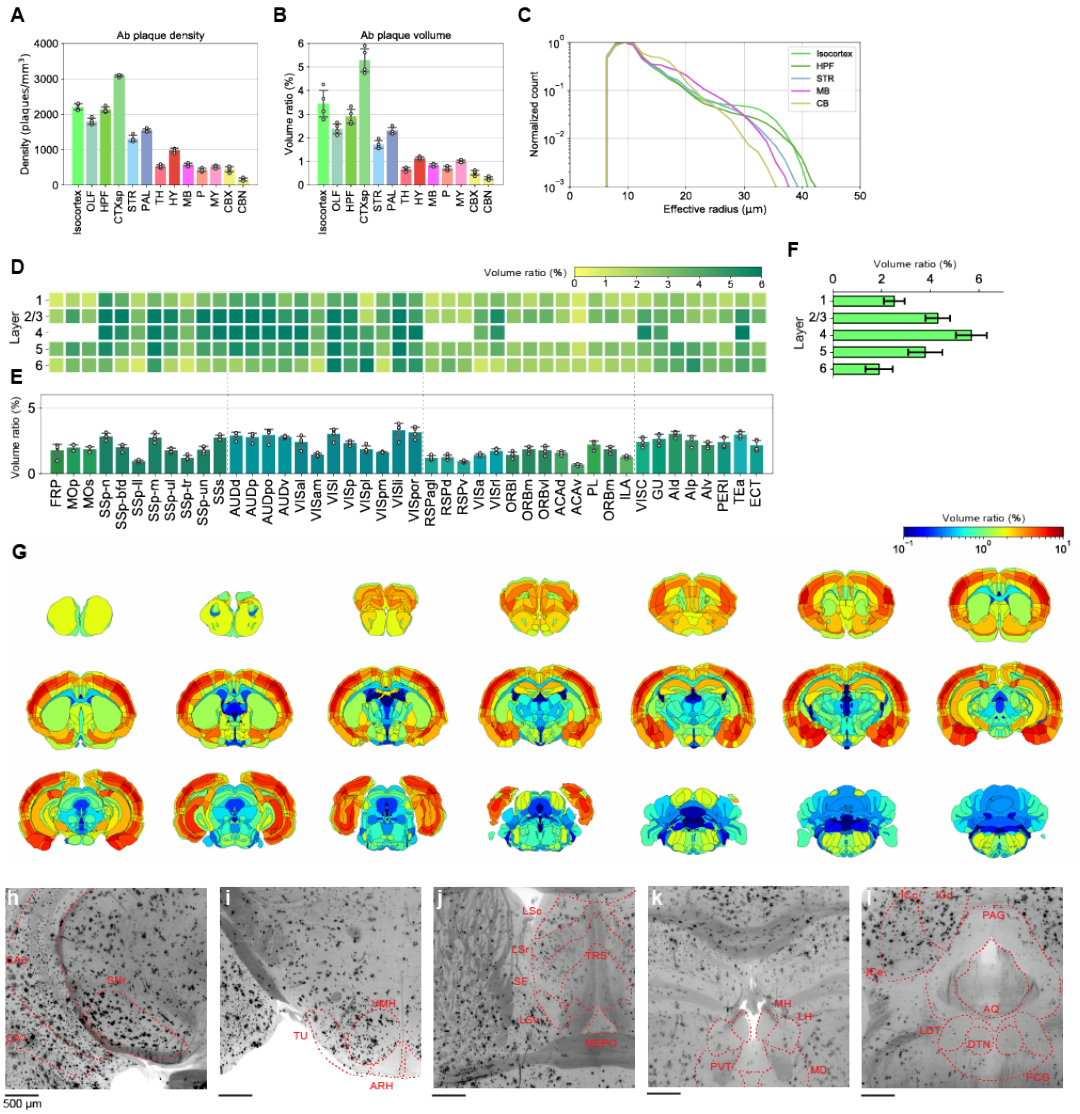

FIGURE 4.13: Whole-brain Analysis of Aβ Plaques Accumulation in AD Model Mouse Brain

A. Density of Aβ plaques (number of plaques/mm^3) in major brain divisions ($n = 4$). **B**. Volume ratio of Aβ plaque in major brain divisions ($n = 4$), computed as (total plaque volume in the region)/(region volume). **C**. Distribution of effective radius of Aβ plaques in the isocortex, hippocampus (HPF), striatum (STR), midbrain (MB) and cerebellum (CB) ($n = 4$). **D,E**. The volume ratio of Aβ plaques in the isocortex ($n = 4$). **F** Layer-wise average of **D**. **G**. Cartoon heatmap showing the Aβ plaque volume ratio in each brain region ($n = 4$). **H-L**. Raw 6E10 immunostaining images around SNr (**H**), VMH (**I**), TRS and MEPO (**J**), LDT and DTN (**K**) and MH, LH and PVT (**L**).

to use the same analysis pipeline of CUBIC-Cloud designed for single-cell analysis. Of note, no plaque staining pattern was observed in the control wild-type mouse brain (9- to 10-month-old, $n = 3$, data not shown).

I first quantified the density (number of individual plaques per volume) and the volume ratio (computed as (total plaque volume in the region)/(region volume)). In both metrics, Aβ plaque amounts were highest in the cerebral cortex and cereberal nuclei, and relatively lower amount of plaques were observed in the brain stem and

cerebellum (Figure 4.13 **A** and **B**). The mean effective radius of the plaque (computed as $r = \{3V/(4\pi)\}^{1/3}$ where V is the plaque volume) tended to be larger in the isocortex and hippocampus and smaller in cerebellum (Figure 4.13 **C**). Within the isocortex, a relatively stronger accumulation of Aβ plaques were observed in visual and auditory areas, whereas plaques were relatively sparse in in medial frontal areas (Figure 4.13 **D** and **E**). Layer-wise abundance of Aβ plaque showed concave profile, with its peak in layer 4 (Figure 4.13 **F**). The whole-brain cartoon heatmap showing the Aβ volume ratio is shown in Figure 4.13 **G**. In the brain stem, the plaque volume ratio was typically 0.5% to 1.0%. Some brain stem regions, however, showed notably larger or smaller amount of Aβ accumulation. For example, SNr and VMH had relatively higher amount of Aβ compared to the neighboring regions (Figure 4.13 **H** and **I**). On the other hand, the ARH, right next to VMH, had almost no Aβ plaques (Figure 4.13 **I**). I also observed that the regions around the ventricles showed relatively lower amount of plaques, including TRS, DTN, MH, LH and PVT (Figure 4.13 **J**, **K** and **L**).

Together, these heterogeneous development of Aβ plaques may reflect the pathological nature of the AD model mouse. Comparison with other AD model mouse lines, such as the recent whole-brain quantification results reported by Liebmann et al., 2016, would be fruitful in elucidating the basic AD pathology.

4.5 Analysis of ARH^{Kiss1+} neuron circuits using Rabies Virus

Contribution statement: The AAV and RV virus production and injection was performed by Dr. E. A. Susaki, Dr. K. Murata and Dr. K. Miyamichi. The brain clearing and staining were done by R. Tanaka. LSFM imaging and data analysis were conducted by the author.

In the previous chapters, I have demonstrated the whole-brain mapping of cell types, IEGs and the disease markers using CUBIC-Cloud. As the last application domain of CUBIC-Cloud, here I report the whole-brain connectivity analysis using pseudotyped rabies virus (RV) (Chapter 1.1.2).

In this experiment, I focused on a population of neurons that secrete kisspeptin (a neuropeptide encoded by *Kiss1* gene) located in the aruate nucleus of hypothalamus (ARH), hereafter termed as ARH^{Kiss1+}. Those neurons were shown to play an important role in reproduction behavior in mammals by regulating pulsatile release of gonadotrophin-releasing hormone (GnRH) at around 0.3 to 1.0 pulses per hour (Herbison, 2018). Intriguingly, the pulse frequency changes through estrus cycle in females but not in males. As such, I investigated the neural inputs to ARH^{Kiss1+} neurons to search the mechanism of pulse generationmodulation on the basis of neural circuitry.

To achieve cell-type specific targeting of virus infection, the Cre/loxP system and RV trans-synaptic tracing combined with Cre-dependent AAV vectors were used (Miyamichi et al., 2013) (Figure 4.14 **A**, Chapter 2.1.5). To identify the sexually dimorphic circuity, injections were performed in both male and female brains. After virus injection, brains were cleared by CUBIC reagents and analyzed by CUBIC-Cloud pipeline (see Chapter 2).

I first checked the localization of the starter cells (GFP+ and mCherry+) to ensure that the injection was successful and the starter cells were well confined within ARH (Figure 4.14 **B**). In the present study, the criteria in selecting successful injection was defined as more than 45% of starter cells were localized in ARH or PVp[2]. With this

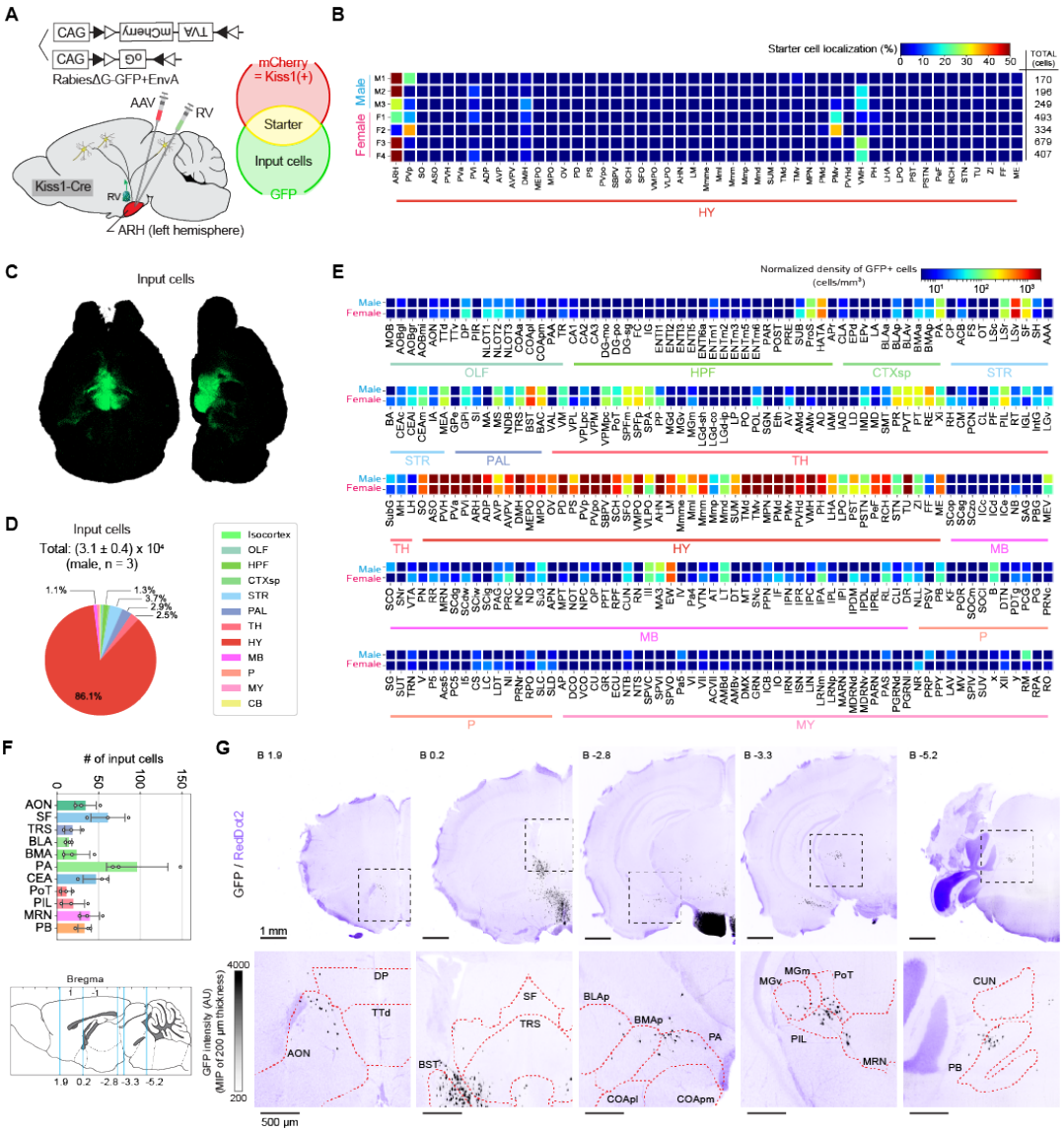

FIGURE 4.14: Whole-brain analysis of input cell populations projecting to ARH^{Kiss1+} neurons

A. Virus injection scheme. AAV carrying mCherry, TVA receptor and optimized glycoprotein (oG) was injected to ARH of Kiss1-Cre transgenic mouse, followed by injection of modified Rabies virus carrying GFP. Cells expressing both mCherry and GFP are the starter cells. **B**. Quantification of starter cell localization. The ratio was computed by dividing the cell count in each region by the total number of starter cells. The total number of starter cells each sample is shown on the right end of the heatmap. **C**. Whole-brain view of all input cells. **D**. Total cell count and the distribution of input cells. Only male brains were considered here. **E**. Cell density heatmap of all brain regions (excluding the isocortex and cerebellum, where no input cells were detected). The mean of male and female brains are shown. **F**. The plot shows extremely sparse input cell populations in previously unidentified brain regions. Only male brains were considered here. **G**. Raw GFP (black) and nuclear staining (RedDot2, purple) images showing the regions identified in **F**. Macro view (top) and zoomed-in view (bottom) are shown. See Table 4.1 for the definitions of brain region acronyms.

criteria, out of 20 injections, $n = 3$ and $n = 4$ brains were assessed as successful for

male and female, respectively (Figure 4.14 **B**).

The whole-brain overview of all input (GFP+ and mCherry-) cells are shown in Figure 4.14 **C**. The present analysis identified $(3.1 \pm 0.5) \times 10^4$ input cells in the male brain ($n = 3$), the majority of which ($> 85\%$) were located within the hypothalamus (Figure 4.14 **D**). As is shown in Figure 4.14 **E**, ARH^{Kiss1+} neurons receive inputs from dozens of discrete structures throughout the forebrain and brainstem, including the striatum (LS), pallidum (BST), thalamus (PVT), hypothalamus (MPO, MPN, AHN, PVH, DMH, VMH and PH), hippocampal formation (HATA and SUB), midbrain (MRN and PAG), and pons (PB). Remarkably, extremely sparse populations, only a few dozens of cells per region, were reproducibly identified (Figure 4.14 **F** and **G**). In terms of cell density, those populations were often equivalent to less than 10 cell/mm^3, which could be easily overlooked with slice-based approaches. These sparse populations were not documented in the past literature studying the ARH^{Kiss1+} neurons (Yeo et al., 2019).

I next performed statistical analysis comparing the number of input cells between male and female brains (Chapter 2.3.1). Overall, binary connectivity differences (i.e. zero in one sex and some finite number in the other sex) were not observed. Weak differences were suggested in LS, MPO, MPN and AVP (Figure4.15), which are neighboring with each other. The difference was most pronounced in LSr, which is known to inhibit the lordosis behavior during mating interactions (Tsukahara, Kanaya, and Yamanouchi, 2014). The sexually dimorphic circuit from LSr to PAG is known, where female brains contain more neurons in LSr that project to PAG (Tsukahara and Yamanouchi, 2002). The present result suggest that LSr sends sexually dimorphic projection to ARH^{Kiss1+}. The identities of these populations can be fully characterized in the future studies.

Of note, using RV injection and slice-based observation, Wang et al., 2015 investigated the input cell population of pro-opiomelanocortin (POMC) neurons and agouti-related peptide (AgRP) neurons in the ARH, another dominant cell types in the ARH. The brain areas containing the input cells to ARH^{Kiss1+} neurons largely overlaped with those of POMC neurons and AgRP neurons. In some areas, however, interesting differences were observed. For example, in VTA and NI, no input cells were detected for ARH^{Kiss1+}, while some input cells were reported to exist for POMC and AgRP. On the other hand, BMA, PA, SF, CEA, SI, TRS, PIL, MRN and PB contained small number of input cells to ARH^{Kiss1+} neurons (Figure 4.14), while they were not reported for POMC and AgRP neurons. This absence of input cells may reflect the actual biological differences, or it may reflect the superior sensitivity of our experimental methods to detect sparse populations.

[2]Note that area annotated as PVp in the Allen Brain Atlas belongs to a part of ARH in the Paxinos atlas (Paxinos and Franklin, 2012).

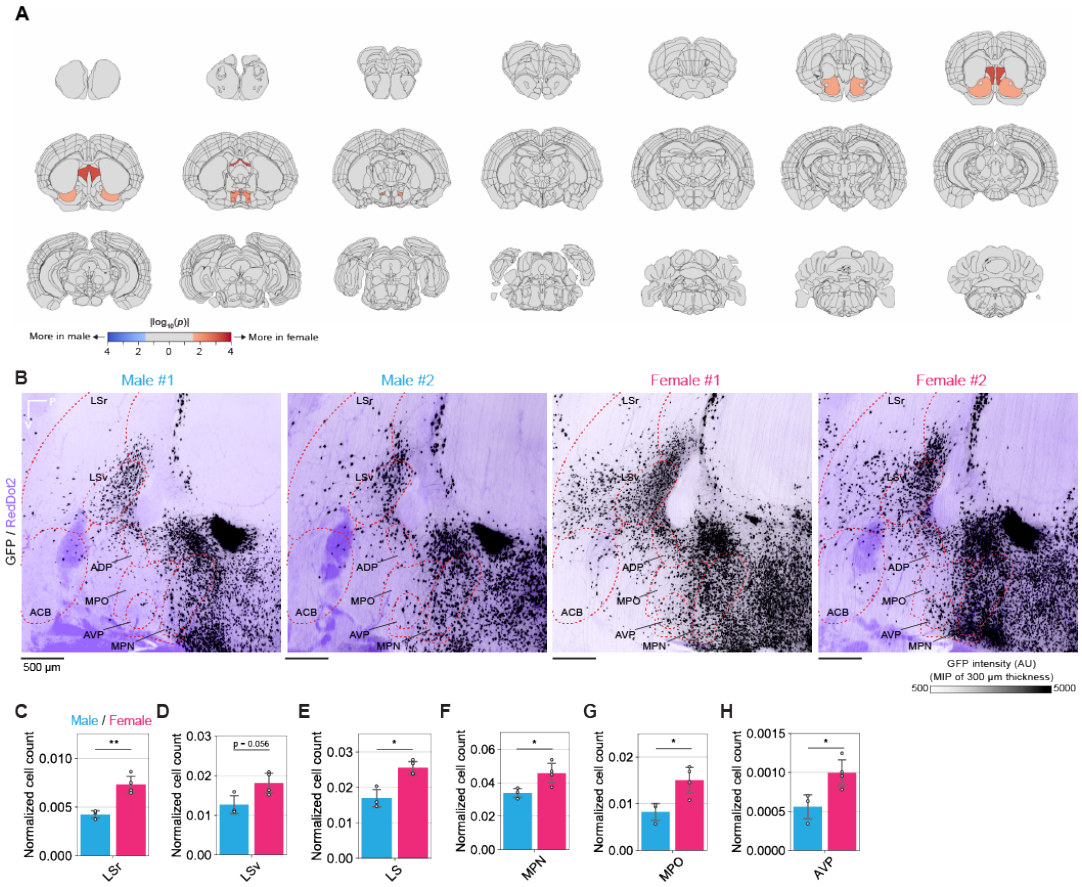

FIGURE 4.15: Sexually dimorphic projection to ARH^{Kiss1+} neurons

A. P-value heatmap where the number of input cells were compared between male and female brains. The color lookup table is log scaled (base 10), where red color represents the regions where more input cells were found in female brains, and blue represents the inverse. Regions with no statistical significance ($p > 0.05$) were assigned a gray color. **B**. Raw GFP (black) and nuclear staining (RedDot2, purple) images around lateral septal nucleus (LS) and medial prepoptic area (MPO). The images are digitally reconstructed sagittal sections. Maximum intensity project (MIP) spanning 300 µm thickness. **C-H**. The plot shows the normalized input cell count in regions where sexual dimorphisms were suggested. *$p < 0.05$, **$p < 0.01$; Welch's t-test. See Table 4.1 for the definitions of brain region acronyms.

TABLE 4.1: Abbreviations of the brain areas

Abbreviations of anatomical structures	
Abbreviation	Full name
AAA	Anterior amygdalar area
ACB	Nucleus accumbens
ADP	Anterodorsal preoptic nucleus
AHN	Anterior hypothalamic nucleus
ARH	Arcuate hypothalamic nucleus
AUDd	Dorsal auditory area
AUDp	Primary auditory area
AUDpo	Posterior auditory area
AUDv	Ventral auditory area
AVP	Anteroventral preoptic nucleus
BST	Bed nuclei of the stria terminalis
CEA	Central amygdalar nucleus
CLI	Central linear nucleus raphe
DMH	Dorsomedial nucleus of the hypothalamus
DMX	Dorsal motor nucleus of the vagus nerve
DTN	Dorsal tegmental nucleus
ECT	Ectorhinal area
FL	Flocculus
HATA	Hippocampo-amygdalar transition area
IA	Intercalated amygdalar nucleus
IC	Inferior colliculus
ILA	Infralimbic area
IO	Inferior olivary complex
LC	Locus ceruleus
LDT	Laterodorsal tegmental nucleus
LH	Lateral habenula
LRN	Lateral reticular nucleus
LS	Lateral septal nucleus
LPO	Lateral preoptic area
MEA	Medial amygdalar nucleus
MH	Medial habenula
MOp	Primary motor area
MOs	Secondary motor area
MRN	Midbrain reticular nucleus
MPN	Medial preoptic nucleus
MPO	Medial preoptic area
NLL	Nucleus of the lateral lemniscus
NOD	Nodulus (X)
NTB	Nucleus of the trapezoid body
NTS	Nucleus of the solitary tract
ORB	Orbital area (ORB)
PAG	Periaqueductal gray
PB	Parabrachial nucleus
PH	Posterior hypothalamic nucleus
PL	Prelimic area
PP	Peripeduncular nucleus
PVH	Paraventricular hypothalamic nucleus

Continuation of Table 4.1

Abbreviation	Full name
PVa	Periventricular hypothalamic nucleus, anterior part
PVi	Periventricular hypothalamic nucleus, intermediate part
PVp	Periventricular hypothalamic nucleus, posterior part
PVpo	Periventricular hypothalamic nucleus, preoptic part
PVT	Paraventricular nucleus of the thalamus
RAmb	Midbrain raphe nuclei
RL	Rostral linear nucleus raphe
RR	Midbrain reticular nucleus, retrorubral area
RSPagl	Retrosplenial area, lateral agranular part
RSPd	Retrosplenial area, dorsal part
RSPv	Retrosplenial area, ventral part
RT	Reticular nucleus of the thalamus
SFO	Subfornical organ
SNc	Substantia nigra, compact part
SNr	Substantia nigra, reticular part
SO	Supraoptic nucleus
SOC	Superior olivary complex
STN	Subthalamic nucleus
SSs	Supplemental somatosensory area
SSp-bfd	Primary somatosensory area, barrel field
SSp-ll	Primary somatosensory area, lower limb
SSp-m	Primary somatosensory area, mouth
SSP-n	Primary somatosensory area, nose
SSp-tr	Primary somatosensory area, trunk
SSp-ul	Primary somatosensory area, upper limb
SSp-un	Primary somatosensory area, unassigned
SUB	Subiculum
TRS	Triangular nucleus of septum
VISal	Anterolateral visual area
VISam	Anteromedial visual area
VISl	Lateral visual area
VISli	Laterointermediate area
VISp	Primary visual area
VISpl	Posterolateral visual area
VISpm	posteromedial visual area
VISpor	Postrhinal area
VMH	Ventromedial hypothalamic nucleus
VTA	Ventral tegmental area
ZI	Zona incerta

End of Table

Chapter 5

Discussion

In this study, I presented an integrated computational framework for single-cell-resolution whole-mouse-brain analysis, named CUBIC-Cloud. The study started with postulating the parallelism between genomics and brain mapping, and I proposed to develop a data integration platform for brain mapping research to promote a genomics-like distributed and collaborative data collection scheme (Chapter 1.4). Inspired by the data integration platforms developed in genomics and in human MRI field, I postulated that the framework should provide (1) a standardization strategy by mapping individual brain to the reference, (2) web-based graphical and programmatic interfaces, (3) a cloud-based toolkit to visualize and quantify the brain data, (4) scalable cloud design for computing and storage and (5) a coherent and organized description of the data records. As I have shown in this study, CUBIC-Cloud addressed these requirements by designing a new software stack embracing the latest cloud technologies, widely available for researchers in neuroscience (Chapter 3).

Further, I demonstrated the usability of the software in a wide range of neuroscientific applications (Chapter 4). Thanks to the scalable cloud system and the standardized analysis routine, this paper successfully analyzed over 120 whole mouse brain, a number which is unheard of in the previous whole brain mapping studies conducted by a single researcher. Thus, the present study illustrates the next generation of the whole mouse brain mapping, with a scalable cloud computing and seamless integration with the open brain repository. Later in this chapter, I will outline the expected future use cases of CUBIC-Cloud (Chapter 5.1).

Nonetheless, the current form of the cloud system should never be considered perfect. As I constructed the system and used the system in real applications, dozens of technical and scientific improvements were discovered. In Chapter 5.2, I will visit these insights obtained through the engagement in this research project, and lay out the future extensions of CUBIC-Cloud.

The whole-brain mapping by tissue clearing and LSFM imaging has just been established in the past few years, and is therefore an emerging technique. The true potential of this novel experimental technique is yet to be seen in the future studies. To conclude this thesis paper, I will postulate some of the important research perspectives in this field, and discuss how CUBIC-Cloud may contribute to such challenges (Chapter 5.3).

5.1 Expected future use cases of CUBIC-Cloud

In chapter 4, I demonstrated the applications of CUBIC-Cloud in three major domains that are of interest in neuroscience research: (1) mapping of the cell-type and the gene expression, (2) reconstruction of the neural activity profile by IEG quantification and (3) neural circuit mapping using rabies virus. Importantly, thanks to the tissue clearing-based 3D imaging and cloud-assisted image analysis, dozens of

whole-brain images can be acquired and analyzed within a few weeks. Therefore, the results presented in this paper will encourage future studies to further accelerate the whole-brain mapping with various labeling targets in a diverse experimental conditions. As the whole-brain mapping gains more momentum in the research field, CUBIC-Cloud may be able to serve as a central hub to integrate the data by many independent studies.

In Chapter 4.1, I showed the whole-brain mapping of five major cell-types. By using whole-brain immunostaining, I quantified the absolute number as well as the expression levels of the PV+, SST+, ChAT+, TH+ and Iba1+ cells, respectively. The data sets presented here establishes a valuable starting point to further explore how the distribution of the cell types are modified under behavioral, environmental, developmental or genetical perturbations. To explore such direction, in this paper I constructed the developmental trajectory of the PV expression of the whole mouse brain (Chapter 4.2). The analysis revealed drastic changes in PV expression in several brain areas, including RT, ZI and APN. Interestingly, RT, ZI and APN, which are spatially separated apart, share similar function that they project to thalamic nuclei and are thought to modulate the sensory processing. The biological meanings of the age-related decrease of PV+ neurons in these areas are to be elucidated in the future studies. Remarkably, the results reported here exemplifies that, using the rich whole-brain dataset, interesting research hypothesis and questions can be generated purely in a data-driven way. This illustrates the unique and remarkable contribution that the the proposed analysis framework can make to the neuroscience research in general.

Built upon the brain-wide cell-type map reported in this study, the future studies would further expand the dataset by targeting different type of cells or performing the experiment under different conditions. For example, autism spectrum disorder (ASD) causes abnormalities in the brain development, and is known to disrupt the expression of various genes in the cortex, including PV. It would be interesting to collect the developmental PV expression map of the ASD model animals, and compare with the healthy mouse's development map reported in this study. The gene expression mapping would also be fruitful in evaluating the efficacy of the drug treatment of these ASD model animals. Such comprehensive analysis of the disease model mouse would not be possible by a single laboratory, and necessitates the collaboration of many researchers. In such situations, CUBIC-Cloud provides a useful platform to exchange the data.

In chapter 4.3, I demonstrated the brain-wide neural activity reconstruction via IEG imaging. As an example case, I performed the injection of LPS, and comprehensively identified brain regions that were activated or repressed by the drug treatment. The same experimental and computational paradigm can be used to study the neural activation under pharmacological or behavioral conditions. If a sufficient number of such data is accumulated in CUBIC-Cloud, I envision that it would enable a new paradigm of brain mapping which allows researchers to construct a model to predict the behavior from the activity profile or predict the activity given some behavioral or environmental input. Such forward and inverse inference problem has been extensively explored in human fMRI studies (see chapter 1.4.4). Single-cell-resolution brain activity mapping could potentially address such brain decoding question with a completely different modality. Already, there are several interesting IEG mapping results in the literature. The study by Renier et al., 2016 identified brain regions that were activated during parental behavior of the female mouse. The recent study by Roy et al., 2019 successfully identified neuron ensembles engaged in the recall of a certain event (called engram cells). It would have substantial impact to

integrate these IEG mapping results to advance IEG-based neural decoding research.

In chapter 4.5, I revealed the brain-wide projection to the ARH^{Kiss1+} neurons using RVΔG tracing technique. With this analysis, I was able to identify extremely sparse populations (less than 10 cell/mm^3) which was not known in the previous literature. In the conventional, literature-based studies, such small population might not have been documented at all, even if researchers were able to discovered them. By uploading the mapping result to CUBIC-Cloud, these minor populations are equally and unequivocally recorded, and other researchers may be able to visit the data later to discover a functional meaning of such small populations. This is where the value of databasing of the RVΔG tracing data is exemplified, and encourages future RVΔG tracing studies to deposit the data in the same way.

5.2 Future extensions of CUBIC-Cloud

5.2.1 Compatibility with other clearing methods

Current implementation of CUBIC-Cloud is optimized for the brain images obtained by the CUBIC clearing method. Although CUBIC is one of the most popularly used clearing method in the field, other clearing methods may be more suitable for certain experimental requirements. In addition, clearing methods are constantly evolving over the years, so new methods with better characteristics may emerge in the future. Further, there are population of researchers who use block-face imaging devices to scan the whole brain (Chapter 1.1.3). To expand the user community of CUBIC-Cloud, the compatibility with other clearing methods needs to be carefully evaluated. It is known that the structural deformation introduced by the clearing method differs depending on the protocol. For example, some of the clearing methods are optimized to preserve the native brain structure (Hama et al., 2015), while hydrophobic clearing methods tend to shrink the brain (Pan et al., 2016) and hydrophilic methods including CUBIC tend to expand the brain. Usually, registration of the brains processed by the same clearing method is easy. Registration of the brains processed by different clearing methods, whose tissue deformation characteristics are significantly different, may be challenging. In the AppendixB.1, I attempted the registration of CUBIC- and iDISCO-cleared brain. The brain registration was successful in many of the brain regions, except for the olfactory bulb where the tissue deformation characteristics were markedly different between CUBIC and iDISCO. To fix this error, I presume that the full-automatic registration methods are not applicable, and requires requires manual annotations on the landmark points in two images.

One possible way to overcome this problem would be to prepare "gateway brains". For example, to register iDISCO cleared brain to the reference brain (cleared by CUBIC), one would first prepare a representative iDISCO cleared brain (called G), and determine the correct transformation between G and reference, possibly involving human corrections. This representative brain (G) is the gateway brain. Then, to map any given iDISCO cleared brains (called X), one would first compute the transformation between X and G. After that, one would apply the two transformation in a sequential manner, to gain a correct mapping from X to the reference. The implementation along this line would be critically important so that the users of CUBIC-Cloud can choose the clearing methods most suited for their experimental purpose.

5.2.2 Cell segmentation in cloud

In the present study, ilastik (Sommer et al., 2011) was used to segment single cells from whole-brain images, computed on the local machines. This design choice was made partly because of the development cost to provide an entirely new cell segmentation program that natively runs in the cloud. However, ilastik consumes substantial computer resources (typically 45 minutes of runtime using 30 CPU cores), and such compute resources may not be readily available at all neuroscience laboratories working on brain mapping. Therefore, to offer ideal usability and accessibility, it is important to migrate cell segmentation task to the cloud so that all computational procedures will be completed within the cloud.

Rather than building a custom cell segmentation software from scratch, it is probably wise to integrate with the existing cloud-based cell segmentation frameworks (Haberl et al., 2018; Bannon et al., 2018; Falk et al., 2019; Wu et al., 2019). In particular, DeepCell 2.0 by (Bannon et al., 2018) proposes an interesting use of cloud for cell segmentation. Their framework not only offer user-friendly graphical interfaces, but also offer means to share the training image datasets, as well as the trained neural networks, with other users. This approach would allow researchers to easily reproduce the image analysis developed by other researchers, or to train a more genelalized neural network by using a massive pool of training data sets provided by the user community. Integration of CUBIC-Cloud with such open and scalable cell segmentation platform may facilitate even more productive collaboration within the user community.

5.2.3 Cloud storage of the raw image data

In the current implementation, CUBIC-Cloud asks the user to submit the segmented cell data to the cloud. However, as demonstrated in human neuroimaging field (Chapter 1.4.4), it is easy to imagine that the data miners may wish to visit the raw image data to extract further information not present in the segmented cell data. CUBIC-Cloud does provide an optional field in the brain data where user can fill the URL to the raw image data. The problem is that in the current form management policy of the raw image data is not precisely defined, and is solely dependent on the user's own action.

In the present study, I have prepared a private cloud server powered by CATMAID (Saalfeld et al., 2009) to provide a web-based interactive viewer to access the raw image data collected in the present study (The server is hosted at `http://cubic-atlas.riken.jp/`). However, setting up and maintaining a private server is not a realistic solution for most of the users, and thus calls for a establishment of the public image data repository. Fortunately, there are several public repository to deposit large biological image data. For instance, IDR (Williams et al., 2017) offers an open platform to allow any researchers to deposit the raw image data. Another collective effort is being advanced by NueroData project (Vogelstein et al., 2018). Relying upon these platforms for raw image storage would be a sensible solution. However, because these platforms are designed for general biological/neuroscientific researches, there may be missing components in terms of functionalities and the ontology set to describe the data. Community-wide involvement and initiatives are thus needed to provide a optimal solution for raw image data storage of the mouse brain mapping.

5.2.4 Mapping of the partial brain data

The core concept of CUBIC-Cloud is that it is optimized and designed for whole-brain imaging datasets. However, there is certainly a value to relax this core concept and accept partial brain data. At the present day, the availability of LSFM devices capable of imaging whole mouse brain is still limited, and many neuroscientists use tissue clearing to scan the partial brain data, such as half hemisphere or thick brain sections. Furthermore, many neuroscience researches focus on particular region of the brain, so the whole brain scan may not be necessary for these researches. Hence, it implies that a significant portion of tissue clearing applications collect partial brain data.

In the literature, many algorithms to map partial brain data to the complete 3D brain space have beeen proposed, often embracing the deep neural networks (Song et al., 2018; Chen et al., 2019). By introducing these methods to CUBIC-Cloud's core functionality, the system can be extended to accept partial brain datasets. In doing so, a careful redesign of the database must be addressed, to maintain the coherence of the deposited data within the database.

5.3 Going beyond CUBIC-Cloud

To conclude this thesis paper, here I will discuss some of the important future research directions of the tissue clearing-based brain imaging and associated informatics and database infrastructures.

5.3.1 Integration with live neural recording modalities

On its own, tissue clearing can reveal only static information of the brain, such as neural connections and gene expressions. For the better understanding of neural computation mechanisms, an effective integration of static structure and dynamical neural firing is indispensable. In particular, two-photon calcium imaging and electrophysiological recordings (including single-cell patch clamp and local field potential measurement by high-density microelectrode array) are two powerful approaches to measure the *in vivo* neural activity. The less explored yet important research direction of tissue clearing would be integrating with these live-recording modalities and accurately mapping the recording result to the reference brain space. Notably, the live neural recording methods and tissue clearing methods are orthogonal experimental method and can readily be combined; that is, after the recording of neural activity, the same brain can be processed by tissue clearing and imaged by LSFM.

There are several benefits of combining neural recording and tissue clearing. First, tissue clearing-based post 3D imaging allows the precise identification of the absolute position of the recorded neurons. In physiology experiments, although the coarse positioning may possible, the exact position of the observed region under two-photon microscope or electrophysiological probe is usually not known. By processing the recorded brain with tissue clearing and scanning the whole brain with LSFM, it becomes possible to identify exactly which cells were recorded at single-cell resolution level. In the case of in vivo imaging, the relative positioning of the recorded neuron may be a good feature to identify the same group of cells in the post-fixed whole-brain scanning image. In the case of electrophysiological recording, one could fluorescently label the probe shank to mark the regions where the probe was inserted through. If such analysis is made possible, one can map

many neural recordings obtained from different experiments/animals to the same reference brain space using whole-brain registration. Furthermore, post-fixed 3D imaging enables to identify the gene expression of the recorded neuron by immunolabeling, further enriching the information. Identification of the projection of the recorded neuron via virus injection would also be possible.

Encouragingly, the combination of high-density microelectrode array recording and tissue clearing has been demonstrated recently (Allen et al., 2019), successfully registering multiple recording sessions to the common atlas space. I expect that the research field will see more and more such applications, where live recorded brain tissue are post-processed by tissue clearing to extract further information and finally mapped to the reference space. In such exciting research avenue, I expect that a data integration platform like CUBIC-Cloud would further increase its value. In this scenario, the data sharing platforms would need to be completely reimagined, to allow researchers to share not only the static structural information, but also the dynamical neural recordings aligned to the reference brain. Despite the software challenges, the value of such database cannot be overstated.

5.3.2 Brain mapping of other organisms

In this paper, I have repeatedly used the analogy between genomics and brain mapping. If this analogy is assumed, it is important to notice that the genome sequencing technology was rapidly applied to various organisms, and the whole genome sequence of thousands of organisms are now available. One can now perform comparative analysis of diverse organisms to obtain valuable insights on the protein functions and structures which diverged or converged across the evolution. Based on this observation, I suggest that the next frontier of the whole-brain imaging technology would lie in advancing the brain mapping of diverse organisms.

Indeed, in the present day, a detailed neuroanatomy is only known in major model organisms, such as nematodes, zebrafish, fruit flies and mouse. In other, often-called non-model organisms, their neuroanatomy has been rarely investigated (Laurent, 2020). Conventionally, the difficulty was the labour and cost to scan the entire brain tissue. Now with the tissue clearing-based whole-brain imaging, this process can be drastically accelerated. Fortunately, because brain tissue composition is chemically common across organisms, the existing tissue clearing methods can be generally applicable to clear the brain of other organisms, such as insects (Pende et al., 2020), marmoset (Susaki et al., 2014), and human (Park et al., 2019; Zhao et al., 2020). Once the brain is cleared, LSFM can be used to rapidly scan the entire brain to survey the gene expression and neural connectivity, in the same way as mouse brain imaging.

If the comprehensive brain map of various organisms become available, researchers can perform comparative analysis across species, just as what is done in genomics. From the comparative analysis, one can, for example, search for common circuit motifs involved in certain information processing conserved across evolution. Other possibility is to survey convergent evolution, where the same functionality (such as vision) is acquired and implemented by completely different circuit structures.

To facilitate the brain mapping of various organisms, the importance of the central hub to integrate such data is once again recognized. The current implementation of the CUBIC-Cloud is solely designed for the analysis of mouse brains. Nonetheless, the analysis pipeline such as cell segmentation and registration would be easily reusable in the analysis of different organisms.

Since the pioneering work by Dodt et al., 2007, the tissue clearing-based whole-brain imaging has witnessed a rapid development during the 2010s. In the next decade, I am excited to see the technique to further evolve and expand its horizon. As the technique is integrated with other live-recording modalities and applied to diverse organisms, the importance of scalable image analysis and data sharing would further increase. Future software infrastructure developments, including ones discussed here, will pave the path toward bottom-up and data-driven elucidation of neuronal functions and circuitry.

Appendix A

Brain Registration Methods

A.1 Brain registration of iDISCO-cleared brain

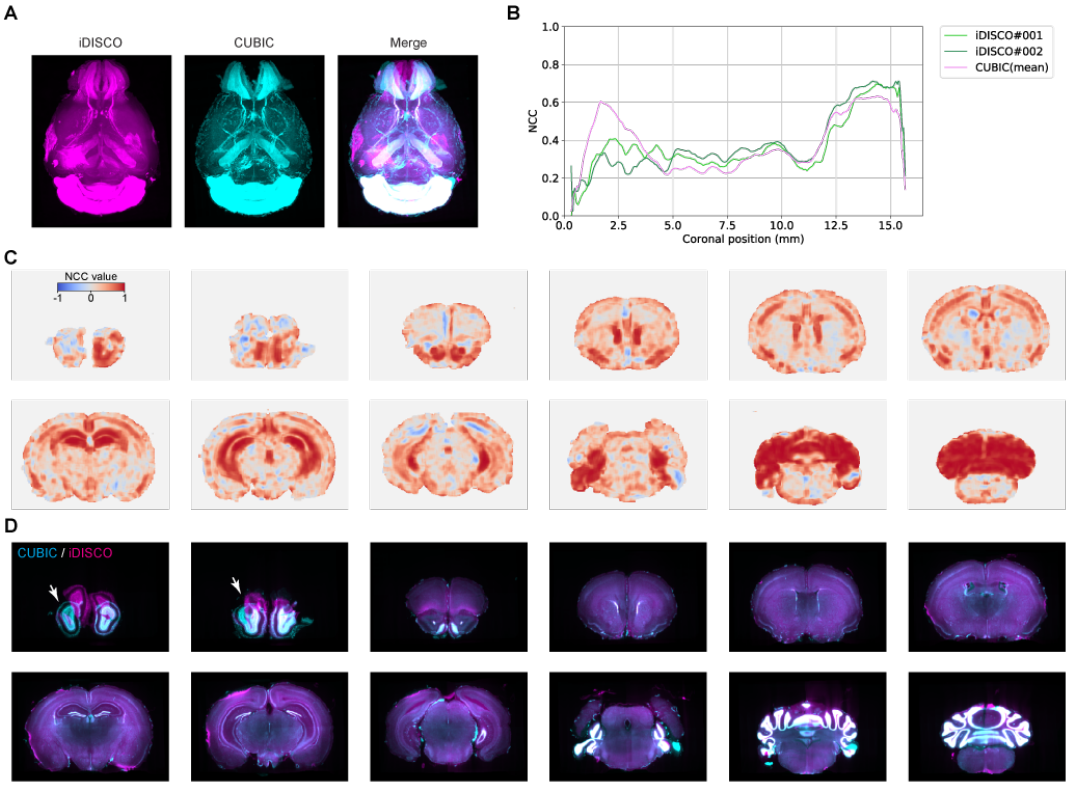

FIGURE A.1: Registration of iDISCO-cleared brain and CUBIC-cleared brain

A. Representative result images of registration between iDISCO- and CUBIC-cleared brains. The brains were stained with nuclear staining dyes. **B**. Normalized cross-correlation (NCC) value between two brains after registration. Mean NCC values of each coronal slice are plotted. Shown in green colors are the results of iDISCO-cleared brains ($n = 2$). Shown in purple is the value of CUBIC-cleared brains (mean of $n = 7$ brains). **C**. Voxel-wise NCC value map computed for the images shown in **D**. **D**. Representative brain registration result. iDISCO (magenta) and CUBIC(cyan) are overlaid. Pointed by arrowheads are the region where the alignment was not accurate.

Different clearing methods cause different deformation to the brain tissue, which makes the brain registration difficult across clearing methods. For example, hydrophobic methods in general shrinks the tissue (Ertürk et al., 2012; Renier et al.,

2014; Pan et al., 2016), whereas hydrophilic methods tend to expand the tissue (Murakami et al., 2018). To ask whether the proposed registration method used in CUBIC-Cloud is compatible with other clearing methods, here I prepared iDISCO-cleared brain and investigated the registration accuracy.

iDISCO (Renier et al., 2014) is organic solvent-based clearing technique and is one of the most popularly used methods due to the high tissue transparency and compatibility with immunostaining. I collected 8-weeks-old C57BL/6N wild-type male mouse brain ($n = 2$) and cleared the tissue following the iDISCO method. The protocol followed the steps described in (Renier et al., 2016), with the following modifications: (1) the immunostainig step was skipped, and (2) the nuclear staining with TO-PRO-3 (ThermoFisher, #T3605) was applied after the tissue permeabilization step. The cleared tissue was imaged with LSFM with (X,Y,Z) = (6.5, 6.5, 7.0) µm voxel resolution. As the imaging oil, the microscope specimen chamber was filled with HIVAC F-4 (Shin-Etsu Silicones), instead of dibenzyl ether (DBE) used in the standard iDISCO method.

The acquired whole-brain nuclear staining image was downscaled to (X,Y,Z) = (30, 30, 30) µm voxel resolution. Then it was aligned with the CUBIC-Atlas using CUBIC-Cloud's registration program with the identical parameter sets as the CUBIC-cleared brain.

The representative registration result is shown in Figure A.1 **A**. A series of coronal slice images are shown in Figure A.1 **D**, along with the corresponding normalized cross-correlation (NCC) values represented as a heatmap (Figure A.1 **C**). The mean of NCC in each coronal slice was plotted in Figure A.1 **B**. As is shown in this plot, the iDISCO-brain registration resulted in similar NCC values as CUBIC-brain registration in most of the brain areas, indicating that the registration method was able to align the iDISCO-cleared brain with the comparable accuracy. However, the registration accuracy was particularly worse in the olfactory bulb (Figure A.1 **A** and **B**). Indeed, the morphology of olfactory bulb is strongly affected in the CUBIC clearing process, where the organ is expanded and a gap between left and right bulb is widened. This misalignment could be corrected by supplying the manually annotated landmarks specifying the corresponding point in the fixed and moving images (Krupa et al., 2020). In conclusion, the current registration method used in CUBIC-Cloud is able to accurately align the iDISCO-cleared brains except for the olfactory areas.

Appendix B

CUBIC-Cloud Documentation

B.1 Step-by-step user guide

In this chapter, a simplified user guide of the CUBIC-Cloud is provided. Users can follow the procedures explained here to start using CUBIC-Cloud. Briefly, the procedure divides into the following steps

- Account setup (Chapter B.1.1)
- Preparing brain data (Chapter B.1.2)
- Uploading brain data (Chapter B.1.3)
- Managing the brain database (Chapter B.1.4)
- Running analysis using notebook (Chapter B.1.5)
- Visualizing brains using studio (Chapter B.1.6)

The GUI design and operations are subjected to the changes in the future updates of the software. The complete and up-to-date documentation of CUBIC-Cloud is available online at `https://cubic-cloud.com/docs/`.

B.1.1 Account setup

The first step to get started with CUBIC-Cloud is creating an account. All users must be logged in with their own account to analyze their own data, as well as to view the published data in the public repository.

To create an account, the user first visit the CUBIC-Cloud's home page. Then, click the *"Sign in"* button at the top-right corner of the web page, or directly go to `https://cubic-cloud.com/signin`. Next, click on a blue button saying *"Create account"* (Figure B.1 **A**). By following the GUI dialogue, users can set the email address and the password of their own account (Figure B.1 **B**).

Once a new account is created, users can log in their account. Once logged in, users are recommended to set the account profile. This information is not mandatory, but it is used when a user publishes data in the public repository. To update the account profile, click *"Account profile"* under the user icon on the top-right corner ((Figure B.1 **C**).

B.1.2 Preparing brain data

In this section I will explain how the user should prepare a whole-brain data to upload and analyze using CUBIC-Cloud. CUBIC-Cloud requires the user to prepare the following data to run analysis

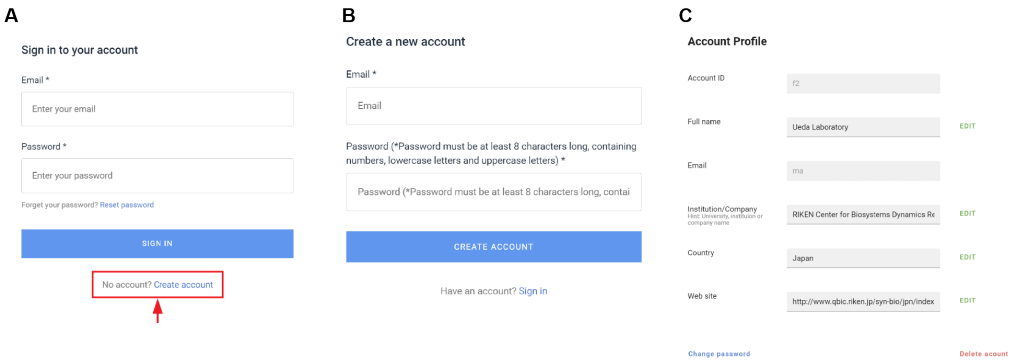

FIGURE B.1: Account setup at CUBIC-Cloud

A, B. A GUI window to create a new user account. **C**. A GUI window to update user profile information.

- **Structure image**: This is a nuclear staining image of the whole mouse brain. This image is used to run brain registration.
- **Cell table**: This is a table which lists the cells in point cloud format extracted from the raw image.

Preparing structure image

To obtain this data, user should stain their brain tissue with the nuclear staining dyes and acquire the whole-brain scan image. So far, the following dyes are tested and validated to work in CUBIC-Cloud.

- RedDot2 (Biotium #40061)
- BOBO-1 (ThermoFisher #B3582)
- SYTOX-G (ThermoFisher #S7020)
- PI (ThermoFisher #P1304MP)

Before uploading to CUBIC-Cloud, users are requested to **rescale the image voxel size to 50 µm**. This operation can be done using any common image analysis software, such as ImageJ (Schindelin et al., 2012). Then the image should be saved in **uncompressed, unsigned 16bit TIFF** format.

Importantly, the 3D image stack should be organized in a **horizontal-major** order, as defined in Figure B.2. This means that the 0th index of the 3D array should correspond to Z, 1st index to Y, and 2nd index to X. If other ordering of the 3D array index is used by the user, it should be rearranged accordingly.

Preparing cell table

As discussed in Chapter 3.1.2, CUBIC-Cloud accepts the whole-brain data in point cloud format. This data should be supplied in a comma-separated values (csv) table format.

To generate the point cloud data, users may use the Python program offered by the current study. The source code and the documentation is available at `https://github.com/DSPsleeporg/ecc`.

Users can choose to use their own cell detection/segmentation program if they wish, as long as the output table follows the format defined here. The table should contain columns named *X*, *Y* and *Z*, which records the center of the mass of the cell in the XYZ coordinate. Here the unit must be **µm**, not voxels. *X*, *Y* and *Z* columns are mandatory, and other columns are optional and can be left blank. However, for example, when a user wants to quantify the expression levels or the cell volume, these columns must be supplied. The column named *deltaI* represents the signal intensity of the cell, after subtraction of the background intensity. The column named *BG* means the background intensity around the cell. The column named *vol* means the volume of the cell, represented in μm^3.

When creating a cell table, the XYZ coordinate ordering must follow that of the CUBIC-Atlas, which is defined as follows (Figure B.2).

- Raw camera image (horizontal section) is viewed from the dorsal side. This means that the anatomical left hemisphere comes to the left-hand side of the image.
- X axis is a left-right axis, where x becomes larger as it goes to right.
- Y axis is an anterior-posterior axis, where y becomes larger as it goes to posterior.
- Z axis is a dorsal-ventral axis, where z becomes larger as it goes to ventral.

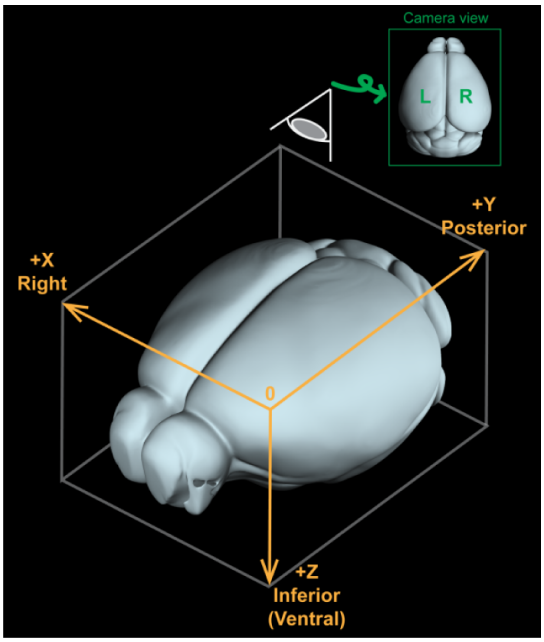

FIGURE B.2: Brain coordinate system used in CUBIC-Cloud

B.1.3 Uploading brain data

In the previous section, I explained the procedures to prepare the data. Next step is to upload this data to CUBIC-Cloud to start the analysis.

To start uploading, click *"Database"* button in the left toolbar, which will show a list of data in the user's private storage space. Then, find a button saying *"New"*

and click it, which will open a new dialogue window (Figure B.3 **B**). In this dialogue, a user can set various attributes of the brain, such as the title, mouse strain name, age and free-text notes. In addition, here user can supply the "Cell labels" tag, which represents the labeling method and targets used. "Experiment label" is used to link different samples collected in a same batch of experiment. After filling these attributes, user would select the structure image and the cell table from the local computer's file system. Once this is done, hit 'Upload' button to start uploading.

After the upload is complete, uploaded data will automatically be placed in a preprocessing job queue. The preprocessing job includes data conversion, registration, and image transformation. In the current version, the preprocessing normally takes about 1 hour. Once preprocessing is complete, users can analyze the data in notebooks or in studios.

B.1.4 Managing the brain database

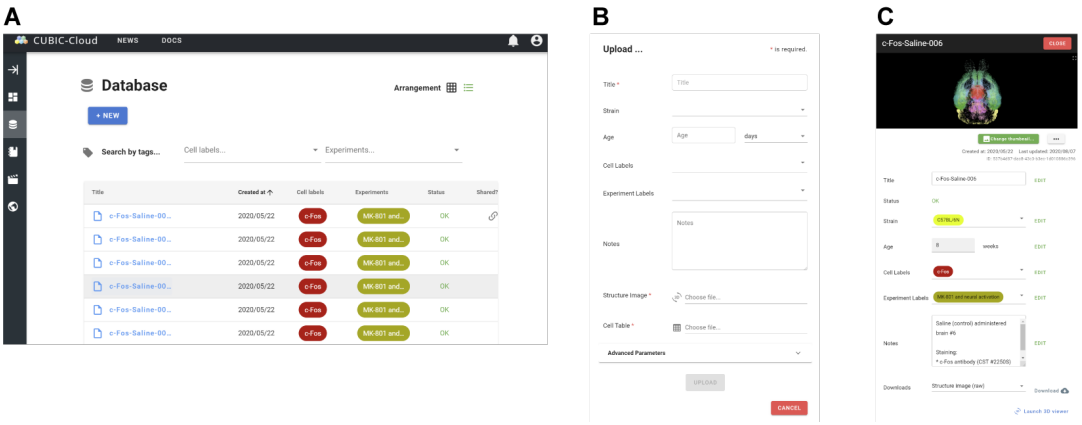

FIGURE B.3: Interfaces for brain database and uploads

A. A GUI window to view the brain data list. **B**. A GUI window to upload a new brain data. **C**. A GUI window to view the detailed information of the brain.

When a user uploads a brain data to the cloud, the data is initially stored in the user's private database. Users can view the list of the brains by clicking *"My list"* button under *"Database"* in the left toolbar (Figure B.3 **A**). The rows can be sorted by the column values, which user can specify by clicking on the header row.

To view the details of the individual item, click on the title of the brain (Figure B.3 **C**). Here, users can view and edit various information attached with the item. In addition, users can download the raw data from this window.

Sharing data

Using share function, a user can let other users access and modify the brain data. The share recipient users can analyze the shared brain data in their notebook, or create a customized visualization using the 3D viewer. To share a brain data, navigate to the *"..."* button at the top right of the brain detail window, and select *"Share"* from the pull-down list. Then, enter the email address of the user, whom you are granting access. Choose either read only or read and write access permission. To remove the user from the shared list, simply click the *"Remove"* button.

Publishing data

Publish allows the user to register a brain data in CUBIC-Cloud public brain repository, where anyone can freely view the data. To publish a brain data, navigate to the *"..."* button at the top right of the brain detail window, and select *"Share"* from the pull-down list. Then click *"Publish"* button. The publish operation cannot be undone, so this operation must be executed with care.

B.1.5 Running analysis using notebook

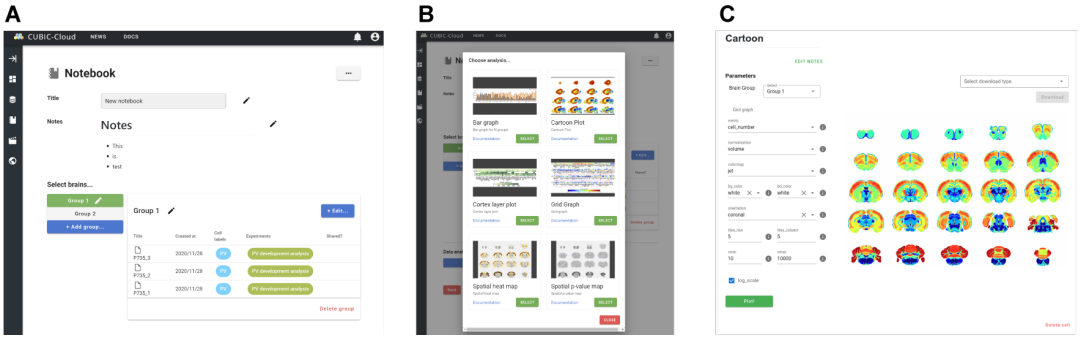

FIGURE B.4: Interfaces for notebook

A. A GUI window to edit the notebook titles and select brains. **B**. A GUI window to select the graph type. **C**. A GUI window to edit the graph parameters.

Using **notebook** users can run diverse whole-brain analysis and generate plots using graphical interfaces. User can access the list of notebooks by clicking on the *"Notebooks"* button in the left toolbar.

To create a new notebook, click *"New"* button in the notebook list page. When a new notebook window opens, first set the title and other additional notes (Figure B.4 **A**). Next, choose the brains to be analyzed (Figure B.4 **A**). Then, click *"Add graph"*, and choose the type of a graph (Figure B.4 **B**). Next, set the parameters required to generate the graph using the text boxes and pulldown menus (Figure B.4 **C**). Finally, hit *"Plot"* button to start generating a graph. Graph generation executed in the cloud server takes between 10 seconds to a several minutes, depending on the type of the graph. Once a graph generation is complete, the generated graph is presented in the GUI (Figure B.4 **C**). The graphs can be downloaded in vector or raster format, as well as in the table format which records the raw numerical values.

Notebooks can be shared and published in the same manner as the brain data. When sharing or publishing a notebook, the brain data included in the notebook is also shared or published, respectively.

B.1.6 Visualizing brains using studio

Using **studios**, users can create a custom visualization of the brain data using interactive 3D brain viewer. Users can access the list of studios by clicking on the *"Studio"* button in the left toolbar.

To create a new studio, click *"New"* button in the studio list page. When a new studio window opens, first set the title and other additional notes. Multiple brains can be virtually overlayed using the studio. Click the *"Add brain"* button, and a pop-up window will show up where user can select the brains from the list (Figure B.5

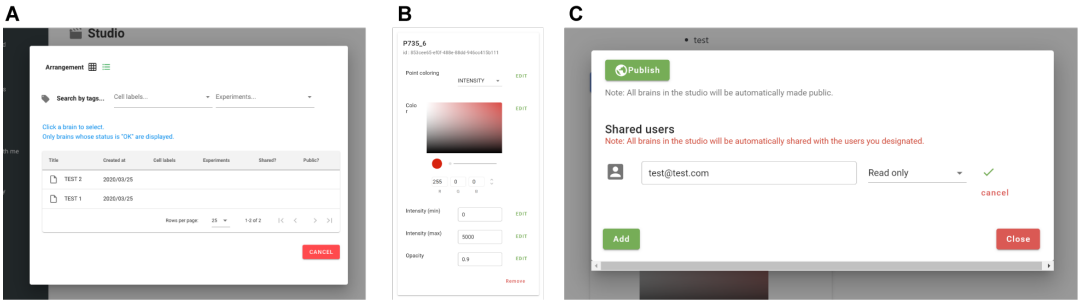

FIGURE B.5: Interfaces for studio

A. A GUI window to select the brains. **B**. A GUI window to set the appearance of the point cloud. **C**. A GUI window to share and publish the studio.

A). The appearance settings (such as color, intensity value range and opacity) can be adjusted for each brain (Figure B.5 **B**). After these parameters are set, click the *"Launch viewer"* button to open a viewer (Figure 3.3).

User can manipulate the viewer using an intuitive mouse click and dragging. Dragging with left click button on the mouse rotates the camera position, while dragging with right click button translates the camera. The mouse wheel controls the zoom. Other interesting visualizations, such as slice views and region-specific views, can be created by following the GUI instructions.

Studios can be shared and published in the same manner as the brain data ((Figure B.5 **C**). When sharing or publishing a studio, the brain data included in the studio is also shared or published, respectively.

Bibliography

Adzic, Gojko and Robert Chatley (2017). "Serverless computing: economic and architectural impact". In: *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering - ESEC/FSE 2017*. New York, New York, USA: ACM Press, pp. 884–889.

Ährlund-Richter, Sofie et al. (2019). "A whole-brain atlas of monosynaptic input targeting four different cell types in the medial prefrontal cortex of the mouse". In: *Nature Neuroscience* 22.4, pp. 657–668.

Allen, William E. et al. (2019). "Thirst regulates motivated behavior through modulation of brainwide neural population dynamics." In: *Science (New York, N.Y.)* 364.6437, p. 253.

Altschul, S F et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." In: *Nucleic acids research* 25.17, pp. 3389–402.

Altschul, Stephen F. et al. (1990). "Basic local alignment search tool". In: *Journal of Molecular Biology* 215.3, pp. 403–410. arXiv: arXiv:1611.08307v1.

Amat, Fernando et al. (2015). "Efficient processing and analysis of large-scale light-sheet microscopy data". In: *Nature Protocols* 10.11, pp. 1679–1696.

Arita, Masanori, Ilene Karsch-Mizrachi, and Guy Cochrane (2020). "The international nucleotide sequence database collaboration". In: *Nucleic Acids Research* 3, pp. 2–5.

Avants, B B et al. (2008). "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain." In: *Medical image analysis* 12.1, pp. 26–41. arXiv: NIHMS150003.

Balakrishnan, Guha et al. (2018). "An Unsupervised Learning Model for Deformable Medical Image Registration". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 9252–9260.

Bannon, Dylan et al. (2018). "DeepCell 2.0: Automated cloud deployment of deep learning models for large-scale cellular image analysis". In: *bioRxiv* 12, p. 505032.

Berg, Stuart et al. (2019). "ilastik: interactive machine learning for (bio)image analysis". In: *Nature Methods* 4522.May, pp. 3–5.

Biswal, Bharat B. et al. (2010). "Toward discovery science of human brain function". In: *Proceedings of the National Academy of Sciences* 107.10, pp. 4734–4739.

Bitzenhofer, Sebastian H, Jastyn A Pöpplau, and Ileana Hanganu-Opatz (2020). "Gamma activity accelerates during prefrontal development". In: *eLife* 9, pp. 1–18.

Boergens, Kevin M. et al. (2017). "webKnossos: efficient online 3D data annotation for connectomics". In: *Nature Methods* 14.7, pp. 691–694.

Cai, Ruiyao et al. (2019). "Panoptic imaging of transparent mice reveals whole-body neuronal projections and skull–meninges connections". In: *Nature Neuroscience* 22.2, pp. 317–327.

Calabrese, Evan et al. (2015). "A Diffusion MRI Tractography Connectome of the Mouse Brain and Comparison with Neuronal Tracer Data". In: *Cerebral Cortex* 25.11, pp. 4628–4637.

Callaway, Edward M. and Liqun Luo (2015). "Monosynaptic Circuit Tracing with Glycoprotein-Deleted Rabies Viruses". In: *Journal of Neuroscience* 35.24, pp. 8979–8985.

Campbell, John N. et al. (2017). "A molecular census of arcuate hypothalamus and median eminence cell types". In: *Nature Neuroscience* 20.3, pp. 484–496.

Chakraborty, Tonmoy et al. (2019). "Light-sheet microscopy of cleared tissues with isotropic, subcellular resolution". In: *Nature Methods* 16.11, pp. 1109–1113.

Chang, Bo-Jui et al. (2019). "Universal light-sheet generation with field synthesis". In: *Nature Methods 2019* 16.March, p. 1.

Chen, Bi-Chang et al. (2014). "Lattice light-sheet microscopy: Imaging molecules to embryos at high spatiotemporal resolution". In: *Science* 346.6208, p. 1257998. arXiv: NIHMS150003.

Chen, Renchao et al. (2017). "Single-Cell RNA-Seq Reveals Hypothalamic Cell Diversity". In: *Cell Reports* 18.13, pp. 3227–3241.

Chen, Yuncong et al. (2019). "An active texture-based digital atlas enables automated mapping of structures and markers across brains". In: *Nature Methods* 16.4, pp. 341–350.

Choi, Dennis C. et al. (2007). "Bed Nucleus of the Stria Terminalis Subregions Differentially Regulate Hypothalamic-Pituitary-Adrenal Axis Activity: Implications for the Integration of Limbic Inputs". In: *Journal of Neuroscience* 27.8, pp. 2025–2034.

Chung, Kwanghun et al. (2013). "Structural and molecular interrogation of intact biological systems". In: *Nature* 497.7449, pp. 332–337. arXiv: NIHMS150003.

Dahlström, Annica and Kjell Fuxe (1964). "EVIDENCE FOR THE EXISTENCE OF MONOAMINE-CONTAINING NEURONS IN THE CENTRAL NERVOUS SYSTEM. I. DEMONSTRATION OF MONOAMINES IN THE CELL BODIES OF BRAIN STEM NEURONS." In: *Acta physiologica Scandinavica. Supplementum* 39.6, SUPPL 232:1–55.

Dean, Kevin M. et al. (2015). "Deconvolution-free Subcellular Imaging with Axially Swept Light Sheet Microscopy". In: *Biophysical Journal* 108.12, pp. 2807–2815.

Dickson, James, Heather Drury, and David C. Van Essen (2001). "'The surface management system' (SuMS) database: a surface–based database to aid cortical surface reconstruction, visualization and analysis". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 356.1412. Ed. by R. Kötter, pp. 1277–1292.

Dodt, Hans-Ulrich et al. (2007). "Ultramicroscopy: three-dimensional visualization of neuronal networks in the whole mouse brain". In: *Nature Methods* 4.4, pp. 331–336.

Donato, Flavio, Santiago Belluco Rompani, and Pico Caroni (2013). "Parvalbumin-expressing basket-cell network plasticity induced by experience regulates adult learning". In: *Nature* 504.7479, pp. 272–276. arXiv: arXiv:1011.1669v3.

Dong, Hong-Wei (2008). *Allen Reference Atlas. A Digital Color Brain Atlas of the C57BL/6J Male Mouse*.

Dorkenwald, Sven et al. (2020). "FlyWire: Online community for whole-brain connectomics". In: *bioRxiv*.

Economo, Michael N. et al. (2016). "A platform for brain-wide imaging and reconstruction of individual neurons". In: *eLife* 5.JANUARY2016, pp. 1–22.

Erö, Csaba et al. (2018). "A Cell Atlas for the Mouse Brain". In: *Frontiers in Neuroinformatics* 12.November, pp. 1–16.

Ertürk, Ali et al. (2012). "Three-dimensional imaging of solvent-cleared organs using 3DISCO". In: *Nature Protocols* 7.11, pp. 1983–1995.

Everitt, Barry J. and Trevor W. Robbins (1997). "Central cholinergic systems and cognition." In: *Annual review of psychology* 48, pp. 649–84.

Falk, Thorsten et al. (2019). "U-Net: deep learning for cell counting, detection, and morphometry". In: *Nature Methods* 16.1, pp. 67–70.

Fürth, Daniel et al. (2018). "An interactive framework for whole-brain maps at cellular resolution". In: *Nature Neuroscience* 21.1, pp. 139–149.

Furube, Eriko et al. (2018). "Brain Region-dependent Heterogeneity and Dose-dependent Difference in Transient Microglia Population Increase during Lipopolysaccharide-induced Inflammation". In: *Scientific Reports* 8.1, pp. 1–15.

Gao, Claire et al. (2020). "Two genetically, anatomically and functionally distinct cell types segregate across anteroposterior axis of paraventricular thalamus". In: *Nature Neuroscience* 23.2, pp. 217–228.

Gao, Liang (2015). "Extend the field of view of selective plan illumination microscopy by tiling the excitation light sheet". In: *Optics Express* 23.5, p. 6102.

Gao, Liang et al. (2014). "3D live fluorescence imaging of cellular dynamics using Bessel beam plane illumination microscopy." In: *Nature protocols* 9.5, pp. 1083–101.

Gao, Ruixuan et al. (2019). "Cortical column and whole-brain imaging with molecular contrast and nanoscale resolution". In: *Science* 363.6424, eaau8302.

Giber, Kristóf et al. (2008). "Heterogeneous output pathways link the anterior pretectal nucleus with the zona incerta and the thalamus in rat." In: *The Journal of comparative neurology* 506.1, pp. 122–40.

Glasser, Matthew F. et al. (2016). "A multi-modal parcellation of human cerebral cortex". In: *Nature* 536.7615, pp. 171–178.

Gong, Hui et al. (2016). "High-throughput dual-colour precision imaging for brain-wide connectome with cytoarchitectonic landmarks at the cellular level". In: *Nature Communications* 7.1, p. 12142.

Gorgolewski, Krzysztof J. et al. (2016). "NeuroVault.org: A repository for sharing unthresholded statistical maps, parcellations, and atlases of the human brain". In: *NeuroImage* 124, pp. 1242–1244.

Gradinaru, Viviana et al. (2018). "Hydrogel-Tissue Chemistry: Principles and Applications." In: *Annual review of biophysics* 47, pp. 355–376.

Greenbaum, Alon et al. (2017). "Bone CLARITY: Clearing, imaging, and computational analysis of osteoprogenitors within intact bone marrow". In: *Science Translational Medicine* 9.387, eaah6518.

Haberl, Matthias G. et al. (2018). "CDeep3M—Plug-and-Play cloud-based deep learning for image segmentation". In: *Nature Methods* 15.9, pp. 677–680.

Hama, Hiroshi et al. (2011). "Scale: a chemical approach for fluorescence imaging and reconstruction of transparent mouse brain". In: *Nature Neuroscience* 14.11, pp. 1481–1488. arXiv: NIHMS150003.

Hama, Hiroshi et al. (2015). "ScaleS: an optical clearing palette for biological imaging". In: *Nature Neuroscience* 18.10, pp. 1518–1529.

Han, Yunyun et al. (2018). "The logic of single-cell projections from visual cortex". In: *Nature* 556.7699, pp. 51–56.

Harris, Julie A. et al. (2019). "Hierarchical organization of cortical and thalamic connectivity". In: *Nature* 575.7781, pp. 195–202.

Herbison, Allan E. (2018). "The Gonadotropin-Releasing Hormone Pulse Generator". In: *Endocrinology* 159.11, pp. 3723–3736.

Hodge, Michael R. et al. (2016). "ConnectomeDB—Sharing human brain connectivity data". In: *NeuroImage* 124, pp. 1102–1107.

Hsu, David T. et al. (1998). "Rapid stress-induced elevations in corticotropin-releasing hormone mRNA in rat central amygdala nucleus and hypothalamic paraventricular nucleus: An in situ hybridization analysis". In: *Brain Research* 788.1-2, pp. 305–310.

Huisken, Jan et al. (2004). "Optical sectioning deep inside live embryos by selective plane illumination microscopy." In: *Science (New York, N.Y.)* 305.5686, pp. 1007–9.

Hunnicutt, Barbara J. et al. (2014). "A comprehensive thalamocortical projection map at the mesoscopic level". In: *Nature Neuroscience* 17.9, pp. 1276–1285.

Inoue, Masafumi et al. (2019). "Rapid chemical clearing of white matter in the post-mortem human brain by 1,2-hexanediol delipidation". In: *Bioorganic & Medicinal Chemistry Letters* 29.15, pp. 1886–1890.

Ito, Daisuke et al. (1998). "Microglia-specific localisation of a novel calcium binding protein, Iba1." In: *Brain research. Molecular brain research* 57.1, pp. 1–9.

Jing, Dian et al. (2018). "Tissue clearing of both hard and soft tissue organs with the PEGASOS method". In: *Cell Research* 28.8, pp. 803–818.

Karolchik, Donna, Angie S Hinrichs, and W James Kent (2009). "The UCSC Genome Browser." In: *Current protocols in bioinformatics* Chapter 1, Unit1.4.

Kasthuri, Narayanan et al. (2015). "Saturated Reconstruction of a Volume of Neocortex". In: *Cell* 162.3, pp. 648–661.

Katz, William T. and Stephen M. Plaza (2019). "DVID: Distributed Versioned Image-Oriented Dataservice". In: *Frontiers in Neural Circuits* 13.February, pp. 1–13.

Ke, Meng-Tsen, Satoshi Fujimoto, and Takeshi Imai (2013). "SeeDB: a simple and morphology-preserving optical clearing agent for neuronal circuit reconstruction". In: *Nature Neuroscience* 16.8, pp. 1154–1161. arXiv: arXiv:1011.1669v3.

Keller, Philipp J et al. (2008). "Reconstruction of Zebrafish Early Embryonic Development by Scanned Light Sheet Microscopy". In: *Science* 322.5904, pp. 1065–1069.

Kim, Euiseok J. et al. (2016). "Improved Monosynaptic Neural Circuit Tracing Using Engineered Rabies Virus Glycoproteins". In: *Cell Reports* 15.4, pp. 692–699.

Kim, Sung-Yon et al. (2015a). "Stochastic electrotransport selectively enhances the transport of highly electromobile molecules". In: *Proceedings of the National Academy of Sciences* 112.46, E6274–E6283.

Kim, Yongsoo et al. (2015b). "Mapping social behavior-induced brain activation at cellular resolution in the mouse". In: *Cell Reports* 10.2, pp. 292–305. arXiv: NIHMS150003.

Kim, Yongsoo et al. (2017). "Brain-wide Maps Reveal Stereotyped Cell-Type-Based Cortical Architecture and Subcortical Sexual Dimorphism". In: *Cell* 171.2, 456–469.e22.

Kirst, Christoph et al. (2020). "Mapping the Fine-Scale Organization and Plasticity of the Brain Vasculature." In: *Cell* 180.4, 780–795.e25.

Klein, Arno et al. (2009). "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration". In: *NeuroImage* 46.3, pp. 786–802.

Klein, Stefan et al. (2010). "elastix: A Toolbox for Intensity-Based Medical Image Registration". In: *IEEE Transactions on Medical Imaging* 29.1, pp. 196–205.

Kleissas, Dean et al. (2019). "The Block Object Storage Service (bossDB): A Cloud-Native Approach for Petascale Neuroscience Discovery". In: *bioRxiv*.

Krupa, Oleh et al. (2020). "NuMorph: tools for cellular phenotyping in tissue cleared whole brain images". In: *bioRxiv*, p. 2020.09.11.293399.

Ku, Taeyun et al. (2016). "Multiplexed and scalable super-resolution imaging of three-dimensional protein localization in size-adjustable tissues". In: *Nature Biotechnology* 34.9, pp. 973–981.

Ku, Taeyun et al. (2020). "Elasticizing tissues for reversible shape transformation and accelerated molecular labeling". In: *Nature Methods* 17.6, pp. 609–613.

Lander, E S et al. (2001). "Initial sequencing and analysis of the human genome." In: *Nature* 409.6822, pp. 860–921.

Lau, Christopher et al. (2008). "Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain." In: *BMC bioinformatics* 9.1, p. 153.

Laurent, Gilles (2020). "On the value of model diversity in neuroscience". In: *Nature Reviews Neuroscience* 21.August.

Lein, Ed S et al. (2007). "Genome-wide atlas of gene expression in the adult mouse brain". In: *Nature* 445.7124, pp. 168–176.

Lewis, David A. et al. (2012). "Cortical parvalbumin interneurons and cognitive dysfunction in schizophrenia." In: *Trends in neurosciences* 35.1, pp. 57–67.

Li, Xiangning et al. (2018). "Generation of a whole-brain atlas for the cholinergic system and mesoscopic projectome analysis of basal forebrain cholinergic neurons." In: *Proceedings of the National Academy of Sciences of the United States of America* 115.2, pp. 415–420.

Li, Yinqing et al. (2020). "Distinct subnetworks of the thalamic reticular nucleus." In: *Nature* 583.7818, pp. 819–824.

Liebmann, Thomas et al. (2016). "Three-Dimensional Study of Alzheimer's Disease Hallmarks Using the iDISCO Clearing Method". In: *Cell Reports* 16.4, pp. 1138–1152.

Lin, Hui-Min et al. (2018). "Reconstruction of Intratelencephalic Neurons in the Mouse Secondary Motor Cortex Reveals the Diverse Projection Patterns of Single Neurons". In: *Frontiers in Neuroanatomy* 12.October, pp. 1–11.

Matsumoto, Katsuhiko et al. (2019). "Advanced CUBIC tissue clearing for whole-organ cell profiling". In: *Nature Protocols* 14.12, pp. 3506–3537.

Menegas, William et al. (2015). "Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass." In: *eLife* 4.AUGUST2015, e10032.

Mennes, Maarten et al. (2013). "Making data sharing work: The FCP/INDI experience". In: *NeuroImage* 82, pp. 683–691.

Miyamichi, Kazunari et al. (2013). "Dissecting Local Circuits: Parvalbumin Interneurons Underlie Broad Feedback Control of Olfactory Bulb Output". In: *Neuron* 80.5, pp. 1232–1245.

Miyawaki, Takeyuki et al. (2020). "Visualization and molecular characterization of whole-brain vascular networks with capillary resolution". In: *Nature Communications* 11.1, p. 1104.

Mok, Tony C.W. and Albert C.S. Chung (2020). "Fast Symmetric Diffeomorphic Image Registration with Convolutional Neural Networks". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4643–4652. arXiv: 2003.09514.

Motta, Alessandro et al. (2019). "Dense connectomic reconstruction in layer 4 of the somatosensory cortex". In: *Science* 366.6469, eaay3134.

Mufson, Elliott J et al. (2008). "Cholinergic system during the progression of Alzheimer's disease: therapeutic implications." In: *Expert review of neurotherapeutics* 8.11, pp. 1703–18.

Murakami, Tatsuya C et al. (2018). "A three-dimensional single-cell-resolution whole-brain atlas using CUBIC-X expansion microscopy and tissue clearing". In: *Nature Neuroscience* 21.4, pp. 625–637.

Murray, Evan et al. (2015). "Simple, Scalable Proteomic Imaging for High-Dimensional Profiling of Intact Systems". In: *Cell* 163.6, pp. 1500–1514.

Nazib, Abdullah, Clinton Fookes, and Dimitri Perrin (2018). "A Comparative Analysis of Registration Tools: Traditional vs Deep Learning Approach on High Resolution Tissue Cleared Data". In: *arXiv*. arXiv: 1810.08315.

Ng, Lydia et al. (2009). "An anatomic gene expression atlas of the adult mouse brain". In: *Nature Neuroscience* 12.3, pp. 356–362.

Oh, Seung Wook et al. (2014). "A mesoscale connectome of the mouse brain". In: *Nature* 508.7495, pp. 207–214.

Ortiz, Cantin et al. (2020). "Molecular atlas of the adult mouse brain". In: *Science Advances* 6.26, eabb3446.

Osakada, Fumitaka and Edward M. Callaway (2013). "Design and generation of recombinant rabies virus vectors". In: *Nature Protocols* 8.8, pp. 1583–1601.

Pan, Chenchen et al. (2016). "Shrinkage-mediated imaging of entire organs and organisms using uDISCO". In: *Nature Methods* 13.10, pp. 859–867.

Pan, Chenchen et al. (2019). "Deep Learning Reveals Cancer Metastasis and Therapeutic Antibody Targeting in the Entire Body". In: *Cell* 179.7, 1661–1676.e19.

Park, Young-gyun et al. (2019). "Protection of tissue physicochemical properties using polyfunctional crosslinkers". In: *Nature Biotechnology* 37.1, pp. 73–83.

Pawley, James B., ed. (2006). *Handbook Of Biological Confocal Microscopy*. Boston, MA: Springer US.

Paxinos, George and Keith Franklin (2012). *Paxinos and Franklin's the Mouse Brain in Stereotaxic Coordinates 4th Edition*. Academic Press.

Pende, Marko et al. (2020). "A versatile depigmentation, clearing, and labeling method for exploring nervous system diversity". In: *Science Advances* 6.22, eaba0365.

Poldrack, Russell A. et al. (2013). "Toward open sharing of task-based fMRI data: the OpenfMRI project". In: *Frontiers in Neuroinformatics* 7.JUNE, pp. 1–12.

Power, Rory M and Jan Huisken (2017). "A guide to light-sheet fluorescence microscopy for multiscale imaging". In: *Nature Methods* 14.4, pp. 360–373.

Qi, Yisong et al. (2019). "FDISCO: Advanced solvent-based clearing method for imaging whole organs". In: *Science Advances* 5.1, eaau8355.

Ragan, Timothy et al. (2012). "Serial two-photon tomography for automated ex vivo mouse brain imaging". In: *Nature Methods* 9.3, pp. 255–258.

Renier, Nicolas et al. (2014). "iDISCO: a simple, rapid method to immunolabel large tissue samples for volume imaging." In: *Cell* 159.4, pp. 896–910.

Renier, Nicolas et al. (2016). "Mapping of Brain Activity by Automated Volume Analysis of Immediate Early Genes". In: *Cell* 165.7, pp. 1789–1802.

Roy, Dheeraj et al. (2019). "Brain-wide mapping of contextual fear memory engram ensembles supports the dispersed engram complex hypothesis". In: *bioRxiv*.

Rueckert, D. et al. (1999). "Nonrigid registration using free-form deformations: application to breast MR images". In: *IEEE Transactions on Medical Imaging* 18.8, pp. 712–721.

Saalfeld, Stephan et al. (2009). "CATMAID: collaborative annotation toolkit for massive amounts of image data." In: *Bioinformatics (Oxford, England)* 25.15, pp. 1984–6.

Sagar, S., F. Sharp, and T. Curran (1988). "Expression of c-fos protein in brain: metabolic mapping at the cellular level". In: *Science* 240.4857, pp. 1328–1331.

Saito, Takashi et al. (2014). "Single App knock-in mouse models of Alzheimer's disease". In: *Nature Neuroscience* 17.5, pp. 661–663. arXiv: NIHMS150003.

Salinas, Casper Bo Gravesen et al. (2018). "Integrated Brain Atlas for Unbiased Mapping of Nervous System Effects Following Liraglutide Treatment". In: *Scientific Reports* 8.1, p. 10310.

Scheffer, Louis K. et al. (2020). "A connectome and analysis of the adult Drosophila central brain". In: *eLife* 9, pp. 1–74.

Schindelin, Johannes et al. (2012). "Fiji: an open-source platform for biological-image analysis." In: *Nature methods* 9.7, pp. 676–82.

Schuetz, Markus (2016). "Potree: Rendering Large Point Clouds in Web Browsers". PhD thesis.

Seiriki, Kaoru et al. (2017). "High-Speed and Scalable Whole-Brain Imaging in Rodents and Primates". In: *Neuron* 94.6, 1085–1100.e6.

Silvestri, L et al. (2012). "Confocal light sheet microscopy: micron-scale neuroanatomy of the entire mouse brain". In: *Optics Express* 20.18, p. 20582.

Sommer, Christoph et al. (2011). "Ilastik: Interactive learning and segmentation toolkit". In: *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, pp. 230–233.

Song, Jun Ho et al. (2018). "Automated 3-D mapping of single neurons in the standard brain atlas using single brain slices". In: *bioRxiv*.

Spalteholz, Werner (1914). *Über das Durchsichtigmachen von menschlichen und tierischen Präparaten und seine theoretischen Bedingungen, nebst Anhang: Über Knochenfärbung*. Ed. by S. Hirzel. Leipzig.

Susaki, Etsuo A. et al. (2014). "Whole-Brain Imaging with Single-Cell Resolution Using Chemical Cocktails and Computational Analysis". In: *Cell* 157.3, pp. 726–739.

Susaki, Etsuo A et al. (2015). "Advanced CUBIC protocols for whole-brain and whole-body clearing and imaging". In: *Nature Protocols* 10.11, pp. 1709–1727.

Susaki, Etsuo A. et al. (2020). "Versatile whole-organ/body staining and imaging based on electrolyte-gel properties of biological tissues". In: *Nature Communications* 11.1, p. 1982.

Tainaka, Kazuki et al. (2014). "Whole-Body Imaging with Single-Cell Resolution by Tissue Decolorization". In: *Cell* 159.4, pp. 911–924.

Tainaka, Kazuki et al. (2016). "Chemical Principles in Tissue Clearing and Staining Protocols for Whole-Body Cell Profiling". In: *Annual Review of Cell and Developmental Biology* 32.1, pp. 713–741.

Tainaka, Kazuki et al. (2018). "Chemical Landscape for Tissue Clearing Based on Hydrophilic Reagents". In: *Cell Reports* 24.8, 2196–2210.e9.

Takesian, Anne E. and Takao K. Hensch (2013). "Balancing Plasticity/Stability Across Brain Development". In: *Progress in Brain Research*. 1st ed. Vol. 207. Elsevier B.V., pp. 3–34.

Tasic, Bosiljka et al. (2016). "Adult mouse cortical cell taxonomy revealed by single cell transcriptomics". In: *Nature Neuroscience* 19.2, pp. 335–346.

Tasic, Bosiljka et al. (2018). "Shared and distinct transcriptomic cell types across neocortical areas". In: *Nature* 563.7729, pp. 72–78.

Tatsuki, Fumiya et al. (2016). "Involvement of Ca 2+ -Dependent Hyperpolarization in Sleep Duration in Mammals". In: *Neuron* 90.1, pp. 70–85.

Thirion, J.-P. (1996). "Non-rigid matching using demons". In: *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 245–251.

Thompson, Carol L. et al. (2008). "Genomic Anatomy of the Hippocampus". In: *Neuron* 60.6, pp. 1010–1021.

Thompson, Carol L. et al. (2014). "A high-resolution spatiotemporal atlas of gene expression of the developing mouse brain". In: *Neuron* 83.2, pp. 309–323.

Todorov, Mihail Ivilinov et al. (2020). "Machine learning analysis of whole mouse brain vasculature". In: *Nature Methods* 17.4, pp. 442–449.

Tomer, Raju et al. (2014). "Advanced CLARITY for rapid and high-resolution imaging of intact tissues". In: *Nature Protocols* 9.7, pp. 1682–1697.

Tsukahara, Shinji, Moeko Kanaya, and Korehito Yamanouchi (2014). "Neuroanatomy and sex differences of the lordosis-inhibiting system in the lateral septum". In: *Frontiers in Neuroscience* 8.SEP, pp. 1–13.

Tsukahara, Shinji and Korehito Yamanouchi (2002). "Sex Difference in Septal Neurons Projecting Axons to Midbrain Central Gray in Rats: A Combined Double Retrograde Tracing and ER-Immunohistochemical Study". In: *Endocrinology* 143.1, pp. 285–294.

Turner, Jessica A. and Angela R. Laird (2012). "The cognitive paradigm ontology: Design and application". In: *Neuroinformatics* 10.1, pp. 57–66.

Ueda, Hiroki R. et al. (2020). "Tissue clearing and its applications in neuroscience". In: *Nature Reviews Neuroscience* 21.2, pp. 61–79.

Van Essen, David C. et al. (2017). "The Brain Analysis Library of Spatial maps and Atlases (BALSA) database". In: *NeuroImage* 144.7615, pp. 270–274.

Van Horn, John D. et al. (2001). "The Functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large–scale databasing of neuroimaging studies". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 356.1412. Ed. by R. Kötter, pp. 1323–1339.

Venter, J C et al. (2001). "The sequence of the human genome." In: *Science (New York, N.Y.)* 291.5507, pp. 1304–51.

Vercauteren, Tom et al. (2009). "Diffeomorphic demons: Efficient non-parametric image registration". In: *NeuroImage* 45.1, S61–S72.

Vettenburg, Tom et al. (2014). "Light-sheet microscopy using an Airy beam". In: *Nature Methods* 11.5, pp. 541–544.

Vogelstein, Joshua T et al. (2018). "A community-developed open-source computational ecosystem for big neuro data". In: *Nature Methods* 15.11, pp. 846–847.

Voigt, Fabian F. et al. (2019). "The mesoSPIM initiative: open-source light-sheet microscopes for imaging cleared tissue". In: *Nature Methods* 16.11, pp. 1105–1108.

Wan, Yinan, Katie McDole, and Philipp J Keller (2019). "Light-Sheet Microscopy and Its Potential for Understanding Developmental Processes". In: *Annual Review of Cell and Developmental Biology* 35.1, pp. 655–681.

Wang, Daqing et al. (2015). "Whole-brain mapping of the direct inputs and axonal projections of POMC and AgRP neurons". In: *Frontiers in Neuroanatomy* 9.March, pp. 1–17.

Wang, Quanxin et al. (2020). "The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas". In: *Cell* 181.4, 936–953.e20.

Wang, Yong et al. (2019). "A bed nucleus of stria terminalis microcircuit regulating inflammation-associated modulation of feeding". In: *Nature Communications* 10.1, p. 2769.

Watabe-Uchida, Mitsuko et al. (2012). "Whole-Brain Mapping of Direct Inputs to Midbrain Dopamine Neurons". In: *Neuron* 74.5, pp. 858–873.

White, J G et al. (1986). "The structure of the nervous system of the nematode Caenorhabditis elegans". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 314.1165, pp. 1–340.

Wickersham, Ian R. et al. (2007). "Monosynaptic Restriction of Transsynaptic Tracing from Single, Genetically Targeted Neurons". In: *Neuron* 53.5, pp. 639–647.

Williams, Eleanor et al. (2017). "Image Data Resource: A bioimage data integration and publication platform". In: *Nature Methods* 14.8, pp. 775–781.

Winnubst, Johan et al. (2019). "Reconstruction of 1,000 Projection Neurons Reveals New Cell Types and Organization of Long-Range Connectivity in the Mouse Brain." In: *Cell* 179.1, 268–281.e13.

Wu, Jingpeng et al. (2019). "Chunkflow: Distributed Hybrid Cloud Processing of Large 3D Images by Convolutional Nets". In: *Frontiers in Neural Circuits* 13, pp. 1–9. arXiv: 1904.10489.

Yang, Bin et al. (2014). "Single-cell phenotyping within transparent intact tissue through whole-body clearing". In: *Cell* 158.4, pp. 945–958.

Yannick, Schwartz, Thirion Bertrand, and Varoquaux Gael (2014). "Mapping cognitive ontologies to and from the brain". In: *Frontiers in Neuroinformatics* 8, pp. 1–9. arXiv: 1311.3859.

Yarkoni, Tal et al. (2011). "Large-scale automated synthesis of human functional neuroimaging data". In: *Nature Methods* 8.8, pp. 665–670.

Yeo, Shel-Hwa et al. (2019). "Mapping neuronal inputs to Kiss1 neurons in the arcuate nucleus of the mouse". In: *PLOS ONE* 14.3. Ed. by Raul M. Luque, e0213927.

Yun, Dae Hee et al. (2019). "Ultrafast immunostaining of organ-scale tissues for scalable proteomic phenotyping". In: *bioRxiv*, p. 660373.

Yushkevich, Paul A. et al. (2006). "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability". In: *NeuroImage* 31.3, pp. 1116–1128.

Zeisel, Amit et al. (2018). "Molecular Architecture of the Mouse Nervous System." In: *Cell* 174.4, 999–1014.e22.

Zhao, Shan et al. (2020). "Cellular and Molecular Probing of Intact Human Organs". In: *Cell* 180.4, 796–812.e19.

Zheng, Ting et al. (2013). "Visualization of brain circuits using two-photon fluorescence micro-optical sectioning tomography". In: *Optics Express* 21.8, p. 9839.

Zheng, Zhihao et al. (2018). "A Complete Electron Microscopy Volume of the Brain of Adult Drosophila melanogaster". In: *Cell* 174.3, 730–743.e22.

Zingg, Brian et al. (2014). "Neural networks of the mouse neocortex." In: *Cell* 156.5, pp. 1096–111.