A Pipeline for Automated Assessment of Cell Location in 3D Mouse Brain Image Sets

Christian J Niedworok

University College London

PhD Supervisor: Troy Margrie

A thesis submitted for the degree of
Doctor of Philosophy
University College London

December 2016

Declaration

I, Christian J Niedworok, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. Part of the work presented in this thesis has been published in Nature Communications (Niedworok et al., 2016). The publication is annexed to this thesis.

Abstract

Mapping the neuronal connectivity of the mouse brain has long been hampered by the laborious and time-consuming process of slicing, staining and imaging the brain tissue. Recent developments in automated 3D fluorescence microscopy, such as serial two-photon tomography (STP) and light sheet fluorescence microscopy, now allow for automated rapid 3D imaging of a complete mouse brain at cellular resolution. In combination with transsynaptic viral tracers, this paves the way for high-throughput brain mapping studies, which could greatly advance our understanding of the function of the brain. Because transsynaptic tracers label synaptically connected cells, the analysis of these whole-brain scans requires detection of fluorescently labelled cells and anatomical segmentation of the data, which are very labour- and time-intensive manual tasks and prevent high-throughput analysis.

This thesis presents and validates two software tools to automate anatomical segmentation and cell detection in serial two photon (STP) scans of the mouse brain. Automated mouse atlas propagation (aMAP) segments the scans into anatomical regions by matching a 3D reference atlas to the data using affine and free-form image registration. The fast automated cell counting tool (FACCT) then detects fluorescently labelled cells in the dataset with a novel approach of stepwise data reduction combined with object detection using artificial neuronal networks. The tools are optimised for large datasets and are capable of processing a 2.5TB STP scan in under two days. The performance of aMAP and FACCT is evaluated on STP scans from retrograde connectivity tracing experiments using rabies virus injections in the primary visual cortex

Acknowledgements

First of all, I would like to thank my supervisor, Troy Margrie for his excellent support during my years as a PhD student. His patience and willingness to tackle ambitious interdisciplinary challenges were truly extraordinary and his ability to ask the right questions provided invaluable scientific input.

I am greatly indebted to Marc Modat for modifying his MRI image registration software (NiftyReg) to run on our datasets and for being an inexhaustible source of knowledge for all my questions regarding image registration and segmentation. I would also like to thank Eli Gibson for his guidance and advice on artificial neuronal networks and for providing a custom version of the "Caffe" deep learning framework.

I would like to thank all the members of the Margrie lab for their help, but especially

Zara Allardyce, for counting cells and proofreading;

Alexander Brown, for carrying out stereotaxic injections, tissue preparations and imaging brains on the serial two-photon microscope, writing several helpful Matlab tools and providing excellent input on Matlab programming and data analysis

Charly Rousseau, for writing the data processing pipeline for our serial two-photon microscope, managing our server infrastructure and providing invaluable advice on programming

Molly Strom, for producing the viruses, taking care of the mice, carrying out stereotaxic injections, tissue preparations and counting cells;

Mateo Velez-Fort, for carrying out stereotaxic injections, tissue preparations and imaging brains on the serial two-photon microscope

Edward Bracey, Sepiedeh Keshavarzi and Anja Schmaltz for proofreading

I am eternally grateful to my friends and family for their incredible support.

I would very much like to thank my second supervisor, James Nelson. Although he has sadly been taken from us far too early, his warm words and helpful advice will always stay with me.

Finally, I am grateful for the funding I received from the Wellcome Trust and the medical research council.

Table of Contents

Table of figures

Abbreviations

ACA	anterior cingulate area
AHN	anterior hypothalamic nucleus
alv	alveus
aMAP	automated mouse atlas propagation
AN	artificial neuron
ANN	artificial neuronal network
AUD	auditory areas
AUD	auditory areas
BS	brain stem
CA1	cornu ammonis area1
CB	cerebellum
cc	corpus callosum
CCF	Allen common coordinate framework mouse atlas
cing	cingulum bundle
CNU	cerebral nuclei
CTXsp	cortical subplate
DNA	deoxyribonucleic acid
EGFP	enhanced green fluorescent protein
ELD	Euclidean landmark distance
ENTl	entorhinal area, lateral part
ENTm	entorhinal area, medial part, dorsal zone
EnvA	avian sarcoma leukosis virus glycoprotein
FACCT	fast automated cell counting tool
ftr	fiber tracts
GAD	glutamate decarboxylase
GPU	graphics processing unit
HIV	human immunodeficiency virus
HPF	hippocampal formation
HPF(ns)	hippocampal formation, no substructure
int	internal capsule
LD	lateral dorsal nucleus of the thalamus
LGd	dorsal part of the lateral geniculate complex
LHA	lateral hypothalamic area
LP	lateral posterior nucleus of the thalamus
LPO	lateral preoptic area
LSFM	light sheet fluorescence microscopy
MO	somatomotor areas
MRI	magnetic resonance imaging
NMDA	N-methyl-D-aspartate
NMI	normalised mutual information
OLF	olfactory areas
PAR	parasubiculum
PFA	paraformaldehyde

POST	postsubiculum
PRE	presubiculum
PTLp	posterior parietal association areas
rAAV	recombinant adeno-associated virus
RG	rabies glycoprotein
RV	rabies virus
RFP	red fluorescent protein
RSP	retrosplenial area
SBA	shape-based averaging
SS	somatosensory areas
SSE	sum of squared errors
STAPLE	simultaneous truth and performance level estimation
STP	serial two-photon tomography
SUB	subiculum
TH	thalamus
TVA	tumor virus A
VIS	visual areas
VISam5	anteromedial visual area, layer 5
VISl5	lateral visual area, layer 5
VISli5	laterointermediate area, layer 5
VISp	primary visual area, layer 5
VISp2/3	primary visual area, layer 2/3
VISp4	primary visual area, layer 4
VISp6a	primary visual area, layer 6a
VPM	ventral posteromedial nucleus of the thalamus
VS	ventricular systems
ZI	zona inerta

Chapter 1. Introduction

The function of the human brain remains one of the greatest unsolved mysteries of our time. Despite advances in our understanding of general principles, such as synaptic plasticity (Bliss and Collingridge, 1993), regulation of network activity (Marder and Goaillard, 2006) or the higher-level function of several brain areas, we currently cannot even fully explain the functionality of the brains of simpler organisms, such as the rat, mouse or fruit fly.

To gain a deeper understanding of the function of these nervous systems, we will need to methodically characterise the function and connectivity of their networks, ideally down to a single neuron, or even a single ion channel. Since such a characterisation is impossible on human beings for practical and ethical reasons, it requires the use of model organisms. There are a number of animal models in use today, from very simple ones, such as the nematode worm Caenorhabditis elegans, with a stereotypical nervous system of exactly 302 neurons that can be simulated in a computer (Szigeti et al., 2014), to highly advanced organisms such as the macaque, which can be trained to perform complex behavioural tasks (Luck et al., 1997).

Currently, mice are the most popular animal model for scientific research, representing 61% of all animals used in research in the UK according to the 2015 Home Office report[1]. Mice are an ideal model organism because they are relatively easy and inexpensive to breed and allow genetic modification (Gordon and Ruddle, 1981). As a result, many specialised transgenic mouse lines that mimic human illnesses or express marker genes in genetically defined cell types are readily available (http://mousemutant.jax.org/). Finally, mice are intelligent enough to learn a variety of behavioural tasks, allowing to manipulate neuronal circuits with the help of genetic tools and investigate the behavioural impact (Yizhar et al., 2011).

[1] Annual Statistics of Scientific Procedures on Living Animals Great Britain 2015, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/537708/scientific-procedures-living-animals-2015.pdf

1.1 History of Brain Mapping

Efforts to characterise the anatomy and connectivity of the brain to gain a better understanding of its function have a long history, dating back to the pioneering neuroanatomical studies of Ramón y Cajal and Golgi (Cajal, 1894, 1896, 1899; Golgi, 1875, 1886, 1989). Recently, the field of brain anatomy has been experiencing a renaissance in the form of connectomics, coined by Sporns et al. (2005) to describe a "comprehensive structural description of the network of elements and connections forming the human brain" but adopted by others to generally describe the analysis of neuronal connectivity on a large scale using modern techniques (Reid, 2012).

1.2 Mapping on Different Scales

Currently, 3D analysis of neuronal connectivity takes place on three scales. Macroscopic approaches focus on magnetic resonance imaging (MRI), with diffusion MRI to evaluate connectivity and functional MRI to evaluate activity ("functional connectivity", Biswal et al. (1995)). The great advantage of these techniques is that they are non-invasive and can be performed on live specimens. However, MRI provides an indirect measure of connection probability and due to its low resolution (~1mm/voxel) can only capture clusters of activity or large fibre bundles that connect different brain regions (Glasser et al., 2013; Marblestone et al., 2013).

At the other end of the scale is dense reconstruction using en electron microscope (EM). Classically, 3D volumes were generated by cutting and imaging serial sections (Sjostrand, 1958; White et al., 1986), but the recent introduction of serial block-face electron microscopy vastly accelerated the speed at which data can be acquired. Here a sample of brain tissue is scanned using an electron microscope, a slice is automatically removed from the surface of the sample and the process is repeated until the complete sample is imaged (Denk and Horstmann, 2004). The resolution of this technique is high enough to identify every single synapse of a neuron, but the acquisition is currently limited to small volumes of ~300μm^3 (Helmstaedter et al., 2008; Morgan and Lichtman, 2013). However, the main limitation for using this technique to map connectivity is that EM data is difficult to analyse. The images are extremely detailed, showing all cells and many of the compartments within them. While this information can be useful, regions

of interest such as the soma, dendrites, axons and synapses need to be traced manually, which is extremely time-consuming. Despite years of intensive research and development, the output of automated analysis methods still requires extensive manual corrections (Helmstaedter, 2013). One group is experimenting with crowdsourcing this kind of analysis (www.eyewire.org, Kim et al. (2014a)), but creating a complete map of the human brain including every synapse is estimated to take in the order of 10 million years at current acquisition and analysis speeds (Morgan and Lichtman, 2013). Even mapping the brain of a smaller organism, such as the mouse, is not yet feasible using EM, although efforts to map smaller regions of the mouse brain are underway (Kasthuri et al., 2015; Kim et al., 2014a).

Representing a middle ground in terms of resolution and complexity, methods based on light microscopy currently appear to be the most promising approach to map connectivity in rodents. While EM gives a highly detailed picture of all cells in a tissue sample, light microscopy methods usually focus on a subset of cells labelled using techniques such as immunohistological stainings, membrane-bound dyes, beads and expression of marker proteins introduced via transgenes or viral transfections (Cowan, 1998; Katz et al., 1984; Lundh, 1990). Classically, tissues are sliced, stained if necessary and imaged using fluorescence or transmitted light microscopy. Because labelled cells represent the strongest signal in the image and the datasets are orders of magnitude smaller than EM data, their analysis is much faster and simpler.

Recent advancements in automated light microscopy methods (see below) now allow a complete mouse brain to be imaged at a sufficiently high resolution to identify individual cells without user interaction (Gong et al., 2013; Niedworok et al., 2012; Ragan et al., 2012). However, these techniques generate large amounts of data at a rapid pace, shifting the major bottleneck from generating data to analysing it. Therefore, the lack of freely available and reliable automated analysis tools for such data represents a major roadblock on the path to determining the connectivity of the rodent brain.

1.2.1 Neuronal Tracing

The methods available for analysing light microscopy data are mainly dependent on whether connectivity is assessed using morphological reconstruction or transsynaptic labelling. Morphological reconstruction of stained neurons is the oldest technique used

to map neuronal connectivity, dating back to the early studies of Ramón y Cajal (Cajal, 1894). Neuronal tracing in light microscopy data can be used for either highly detailed analysis or large-scale approximation of neuronal connectivity.

In the first case, the labelled cells are reconstructed in 3D, either manually or semi-automatically (Peng et al., 2014), enabling detailed evaluation of their morphology. Although the resolution does not allow the identification of single synapses (typical resolutions used for tracing are around ~500-1000nm/pixel), synapses can be specifically labelled using antibodies and the morphology of the neurites can provide a good estimate of connectivity (Chklovskii, 2004). Although morphological reconstruction of individual neurons can provide a comprehensive picture of a cell's connectivity, the analysis is very time-consuming, making it unfeasible for high-throughput or long-range connectivity tracing.

While BigNeuron, a recently started collaboration, aims to eventually automate this detailed morphological analysis and enable its use in high-throughput studies (Peng et al., 2015), currently a more crude form of neuronal tracing is used. Here, a brain area is bulk-labelled, e.g. via injection of a viral marker such as recombinant adeno-associated virus (rAAV). The virus causes expression of a fluorescent marker protein, and fluorescence signal outside the injected area is used to indicate connectivity with the injected area (Hunnicutt et al., 2014; Oh et al., 2014). This is considered to provide a good approximation of connectivity, however it does not report synaptic connectivity.

1.2.2 Transsynaptic Tracers

To facilitate mapping of functional connectivity, various markers have been developed that can cross the synapse and thus label connected cells. The first examples of such transsynaptic tracers were radioactively labelled precursor molecules (e.g. [^{3}H]proline, [^{3}H]leucine), horseradish peroxidase or fluorescent beads (Bennett et al., 1973; Jones and Hartman, 1978; Katz et al., 1984; Kristensson et al., 1971). These pioneering methods are based on passive transport and markers are thus diluted at every synapse. As a result, the tracers are only able to produce a weak and partial staining of connected cells (Büttner-Ennever et al., 1981; Jankowska, 1985; Jones and Hartman, 1978).

To overcome the issue of marker dilution, neurotropic viruses that can cross the synapse were employed as a new form of active transsynaptic tracers. Examples are herpes

simplex virus (Kristensson et al., 1982), suid herpesvirus (pseudorabies; Martin and Dolivo (1983)), vesicular stomatitis virus (Lundh et al., 1987) and rabies virus (RV; Gillet et al. (1986)). Unlike passive tracers, viral tracers actively replicate after crossing the synapse, leading to a robust expression of proteins in all infected cells (Kuypers and Ugolini, 1990). Depending on the type and strain of virus used, the virus crosses the synapse either strictly anterogradely (Sun et al., 1996), retrogradely (Ugolini, 1995) or in both directions (Lundh, 1990). Initial tracing studies used wild type viruses and relied on antibody stainings to visualise the infected cells, but later versions were based on modified versions of viral tracers that cause neurons to express fluorescent markers (Jansen et al., 1995; Mebatsion et al., 1996), enabling more straightforward and reliable analysis.

RV, in particular, has been the target of several modifications that greatly enhance its usefulness in mapping neuronal connectivity (Luo et al., 2008; Osakada et al., 2011; Rancz et al., 2011; Wall et al., 2010; Wickersham et al., 2007a; Wickersham et al., 2007b). Wickersham et al. (2007b) provided the basis for these modifications by developing a monosynaptic tracing protocol based on a glycoprotein-deficient RV (Etessami et al., 2000) that only labels neurons directly connected to the initially infected neurons. RV uses its capsid glycoprotein (RG) to cross synapses in a strictly retrograde fashion and infect cells that provide input to the initially infected cell (Miyamichi et al., 2011; Osakada et al., 2011; Wickersham et al., 2007a). However, in the modified version, the sequence for RG has been replaced by that of a fluorescent markers such as EGFP or mRFP (Figure 1-1 a) and RG is provided by a non-transsynaptic vector such as rAAV. This vector is injected either prior to (Figure 1-1 b, Wickersham et al. (2007b)) or mixed with RV (Niedworok et al., 2012). Within the initially infected cells, rAAV provides the RV with RG, which is necessary for the assembly of functional virus particles that can cross the synapse (Figure 1-1, b & c). However, as RG is not present in the presynaptic cells, RV cannot spread any further. As a result, the method provides a monosynaptic label of upstream connectivity (Figure 1-1 c).

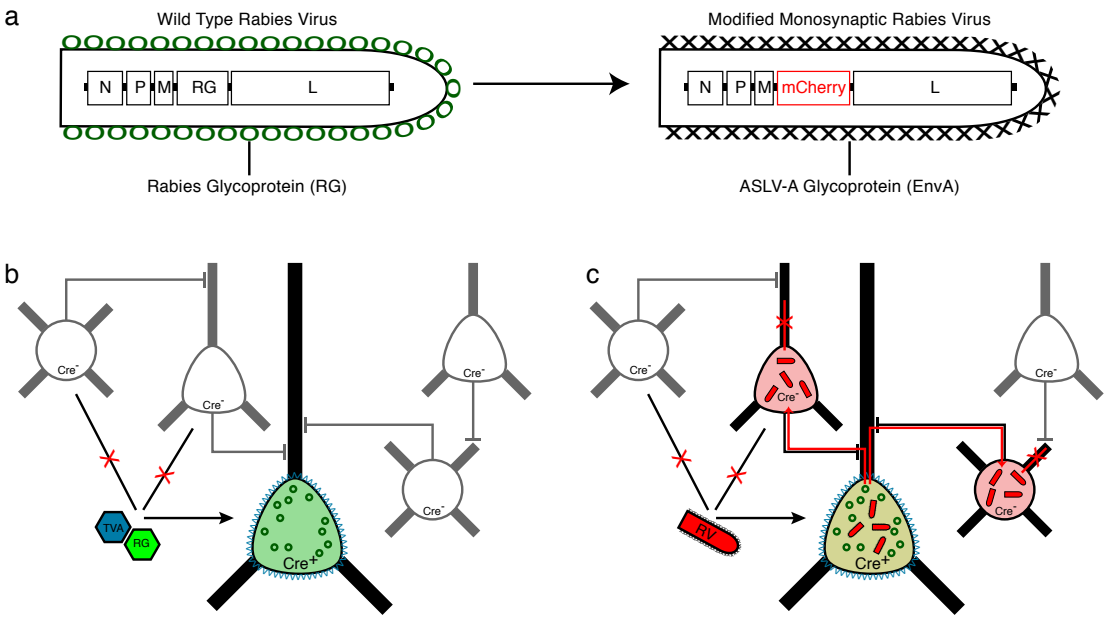


Figure 1-1: Schematic of Cre-dependent monosynaptic rabies tracing

a) Wildtype RV is modified by pseudotyping with EnvA and replacement of its RG gene with a fluorescent protein (mCherry). This renders the RV unable to infect mammalian cells or cross the synapse (b). At the injection site, the RV is complemented by two Cre-dependent rAAVs, one expressing the receptor for EnvA (TVA), the other expressing RG. (d) 11 days later RV is injected at the same location, where it can infect the Cre-positive, rAAV infected cells and cross the synapse. Since the presynaptic cells lack RG, the infection cannot spread further, resulting in monosynaptic labeling. RV: rabies virus; RG: rabies glycoprotein; EnvA: avian sarcoma leukosis virus glycoprotein; TVA: tumor virus A (EnvA receptor); Adapted from Wickersham et al. (2007b)

To increase the specificity of the tracing, the population of "target cells" can be restricted to defined cell types by packaging the RV with the glycoprotein of the avian sarcoma leukosis virus (EnvA), a process called pseudotyping (Figure 1-1 a, Wickersham et al. (2007b)). EnvA lacks an endogenous receptor in mammals (Lewis et al., 2001), which results in an RV that is unable to infect mammalian cells unless they have been modified to express the tumor virus a/EnvA receptor (TVA) (Wickersham et al., 2007b). By using Cre-recombinase-expressing mouse lines in combination with stereotaxic injection of a Cre-dependent rAAV expressing TVA, the initial cell

population can be restricted to genetically and spatially defined cell types (Figure 1-1, b&c; (Lo and Anderson, 2011; Wall et al., 2010)).

This system can also be used to trace the input onto a single cell. This is accomplished by in-vivo electroporation (Marshel et al., 2010) or patch-clamping using an internal solution containing plasmids for expression of RG and TVA (Rancz et al., 2011). RV is then injected at the target location after transfecting the cell, resulting in a labelling of the presynaptic input to the original cell (Marshel et al., 2010; Rancz et al., 2011).

1.3 Advances in Light Microscopy

While transsynaptic tracers remove the time-consuming task of having to trace axons and dendrites to evaluate cell to cell connectivity, the brain tissue still needs to be sliced and imaged. Recent advances in fluorescence microscopy, such as the combination of light sheet fluorescence microscopy (LSFM) with new clearing methods (Chung et al., 2013; Dodt et al., 2007; Renier et al., 2016; Schwarz et al., 2015) or the development of serial two-photon tomography (STP, Ragan et al. (2012)) allow automated generation of high resolution 3D imaging data from large tissue samples, paving the way for high-throughput mapping of the mouse brain.

1.3.1 Light Sheet Fluorescence Microscopy

In LSFM, the brain is illuminated with a thin sheet of light and images are captured with a camera whose light path is perpendicular to the illumination plane. It can therefore be used on transparent samples such as zebrafish embryos and its use with rodent brain tissue requires protocols that clear the specimen. LSFM completely removes the need to slice the sample and thus permits multiple scans using different imaging parameters. In addition, the acquisition time per image is considerably shorter than with scanner-based microscopes, since the whole image is illuminated in a single acquisition frame (50-200ms vs ~1s for an STP microscope).

Unfortunately, successful tissue clearing has been elusive until recently. The first published protocol severely reduced the fluorescence signal and was incompatible with antibody staining procedures (Dodt et al., 2007), while a second protocol failed to achieve the level of transparency necessary for LSFM (Hama et al., 2011). However, several recently published clearing methods have made advances towards solving these

issues and have turned light sheet microscopy into a promising candidate for large-scale brain analysis (Chung et al., 2013; Lee et al., 2016; Renier et al., 2016; Schwarz et al., 2015). However, because the technology is still in its infancy, a limited number of devices can scan samples the size of a mouse brain (Ultramicroscope, LaVision BioTec GmbH; Open SPIM project). So far there are few specialised objectives that combine high optical resolution with large working distances, so an adult mouse brain has to be scanned in at least two orientations (dorsal and ventral) as the effective resolution decreases with increasing depth (Menegas et al., 2015).

Crucially, while the LSFM has a short acquisition time, the brains need to be cleared before scanning, which usually takes several days. Furthermore, clearing protocols can change the morphology by shrinking or enlarging the tissue (Richardson and Lichtman, 2015). If immunohistology is required, some protocols require another week for rehydration, followed by the time needed for slicing and immunohistology (Niedworok et al., 2012). The CLARITY clearing protocol on the other hand requires 6 days of antibody incubation and washing to stain 1mm sections and 6 weeks to stain a complete brain (Chung et al., 2013). However due to its potential for rapid 3D imaging of complete organs, tissue clearing and LSFM are extremely active areas of research.

1.3.2 Serial Two-Photon Tomography

In contrast, STP is an amalgamation and automation of conventional slicing and imaging methods. A two-photon microscope is used to scan the surface of an agarose-embedded tissue sample, which is then automatically transferred to a vibratome, cut to remove a slice and moved back under the microscope. This process is repeated until the whole brain is sliced and imaged at x/y resolutions of up to 500nm/voxel, high enough to visualise single spines if necessary (Ragan et al., 2012). By using a short pulse Ti:Saphire laser, rather than a conventional non-pulsed laser, it is possible to achieve sufficient tissue penetration while minimising photobleaching of out-of-focus areas. The system is adjusted to scan below the surface of the brain to avoid slicing artefacts in the image. The combination of a laser scanning system and physical sectioning of the brain leads to a longer scan time (~60hrs per brain at a resolution of 1µm per pixel and a z-spacing of 5µm) and because the tissue is cut during the process, a brain can only be scanned once. While there is only one commercial system (TissueCyte 1000, Tissue

Vision Cambridge) and one open platform (Economo et al., 2016) available, the optics and mechanics of the system do not differ substantially from well-established microscopy and histology components. Also, as opposed to LSFM, there are no special requirements for preparing and embedding of the tissue. The brain slices could theoretically be stained and re-imaged for further analysis, however this is a time consuming, error-prone manual process (Ragan et al., 2012). While STP can be used in combination with clearing methods (Economo et al., 2016), it is not a prerequisite for the technique.

1.3.3 The Choice of Microscope

The capability to produce high resolution full-brain 3D data, combined with high throughput mean both LSFM and STP are useful for imaging intact brains that have been injected with transsynaptic tracers to map both local and long-range neuronal connectivity. LSFM does not require an expensive two-photon laser system and thus has the potential to eventually become the more widespread method, however at present suboptimal optics and the lack of a well-established scanning system for large samples prevent it from reaching its full potential.

While STP requires longer scanning times, it is not dependent on laborious and time-consuming clearing procedures. The technology is based on well-established components and in our experience, reliable results can be achieved without changing existing experimental protocols. While our lab decided to use STP, and this method is relied upon for this study, the tools developed in this project should be applicable to other types of high throughput whole-brain imaging.

1.4 The Analysis Bottleneck

For a single adult mouse brain imaged in three fluorescent channels, STP generates approximately 2.5TB of single images, each covering a field of view of approximately 1mm^2. These images need to be stored, archived, assembled into coherent 3D datasets (“stitched”), and finally analysed. While handling such amounts of data is a rather new problem in neuroscience, many other sectors such as internet services (Google), particle physics (CERN) and genetics (Human Genome Project) have been working with

similarly large datasets for some time, so the necessary hardware is commercially available.

Since the whole brain is too large to fit into a single field of view, it is mounted on an x-y-stage and scanned using 9x6 partially overlapping images ("tiles") per layer. The first processing step after acquisition corrects for depth-dependent illumination changes, aligns the tiles and stitches them into complete image planes. For this, our lab uses a pipeline combining ImageJ plugins and scripts written in Python. These scripts make heavy use of parallel processing (multithreading) to effectively use the available computational resources and allow to process the STP data in roughly the same time it takes to acquire them. This software pipeline is fully automated and provides a comprehensive 3D volume of the whole mouse brain. However, as with many "big data" projects, this results in a new bottleneck: Classic manual analysis can no longer keep up with the speed at which new data can be generated.

1.4.1 Mapping the Brain

The first crucial step in the analysis of these whole-brain STP datasets is to accurately and reliably determine the locations of anatomical regions in the dataset. Classically, anatomical segmentation has been a manual process with no direct means of quantifying the accuracy of the result. While manual segmentation by expert anatomists is still relevant, especially on 2D data and data suffering from e.g. slicing artefacts, it has significant drawbacks, in particular on 3D fluorescence data: The background fluorescence of brain tissue only provides few anatomical landmarks compared to histological stainings. So while some structures such as the hippocampal formation or cortical regions can be easily identified, the correct segmentation of hypothalamic nuclei or different cortical subregions is much more challenging, even for a skilled neuroanatomist. In addition, variations between animals can complicate the correct assignment of structures, because irregularities in the prominent structures used as reference points can make it more difficult to estimate the correct location and shape of anatomical regions that do not have clear visual boundaries. Irregularities in ventricle size or shape, for example, may lead to choosing the wrong plane from the reference atlas, which will corrupt all segmentations based on that assumption. Different anatomists may also have different opinions about the locations of discrete borders,

which may make it more difficult to meaningfully compare results between different labs. Furthermore, it has been shown on MRI data that a human rater is likely to show significant variation when segmenting the same data multiple times (Nestor et al., 2013). The impact of these issues can be reduced by segmenting the same data multiple times using multiple human raters and generating a "consensus segmentation" from all individual segmentations (Jorge Cardoso et al., 2013; Rohlfing and Maurer, 2007; Warfield et al., 2004). However, manual segmentation is already a very tedious and time-consuming task, and multiple segmentations only exacerbate this issue. Hence, this technique is mostly used in the clinical field, where the correct segmentation of e.g. a tumour is of absolute importance (Warfield et al., 2004).

To improve the speed and reliability of 3D segmentation, automated and semi-automated segmentation algorithms have been pioneered by the MRI community (Collins et al., 1995; Haller et al., 1997), who have been working with 3D datasets since the introduction of clinical MRIs in the 1980s (Mallard, 2003). Automated segmentation has been introduced to the field of rodent fluorescence imaging with the recent emergence of techniques such as the LSFM (Dodt et al., 2007; Renier et al., 2016) and STP (Oh et al., 2014; Ragan et al., 2012) that are capable of imaging a complete rodent brain in 3D at cellular resolution. The segmentation approaches that are most promising for rodent data are based on image registration and atlas propagation. These methods require an atlas, which is comprised of a reference brain (usually an average of multiple 3D brain scans) and a manual anatomical segmentation of this reference brain. The reference brain is then registered to the individual dataset (or vice versa), meaning that it is "deformed" to "best fit" the individual dataset. The manual anatomical segmentation of the reference brain is thus imposed on the individual 3D dataset (Jorge Cardoso et al., 2013; Klein et al., 2010; Ma et al., 2014). The set of available deformations and the definition of a "good fit" are dependent on the registration technique and vary between registration toolkits.

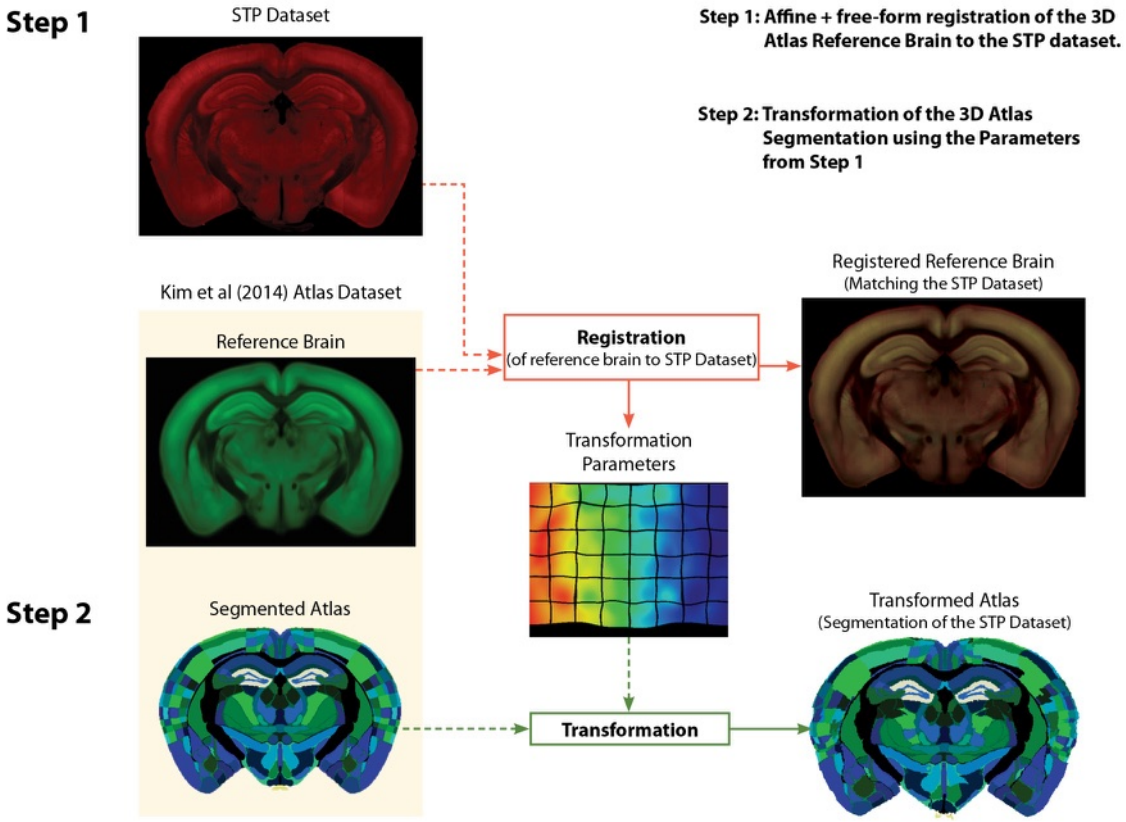


Figure 1-2: Schematic of segmentation propagation

In a first step, the reference brain dataset of the atlas is registered to an individual 3D dataset (here: STP data), using affine and free-form registration. As a result, the registered reference brain now matches the shape of the STP dataset. The transformation parameters that describe the image registration are then applied to the 3D segmentation file that contains the brain structure outlines of the mouse atlas (step 2). As a result, the transformed segmentation file now contains the brain structure outlines describing the anatomy of the STP dataset.

1.4.2 Mouse Atlases

The atlases currently available can be separated into those based on MRI data and those based on serial histological sections (slice-based). While MRI atlases offer full 3D segmentation, they are hampered by the relatively low resolution of MRI data and contain only major brain structures (Johnson et al., 2010; Ma et al., 2005; Ma et al., 2008) or only map certain brain areas in reasonable detail (Richards et al., 2011; Ullmann et al., 2012; Ullmann et al., 2014; Ullmann et al., 2013) making them of limited use for mapping high-resolution STP data. Slice-based atlases on the other hand,

suffer from deformations generated during slicing and handling of the sections, which leads to high frequency noise along the axis perpendicular to the cutting plane (Ng et al. (2007), see 2.7.2), but provide very detailed segmentation of individual structures. Currently, there are two major slice-based atlases. The first is the Paxinos and Franklin (2004), which is based on 120 µm thick coronal sections of one mouse brain, alternately stained using nissl and cresyl violet. These sections were then manually segmented into over 900 brain structures. However, this atlas is only available in printed form. While recent versions also include a digital PDF version of the book, extracting the anatomical structures from this format is difficult and the atlas still lacks a suitable reference-brain dataset for 3D fluorescence images. The second major atlas is the Allen brain atlas (Lein et al., 2007; Ng et al., 2007). It was created using a series of 25µm thick nissl-stained coronal sections that were manually segmented into "several hundred" structures (Ng et al., 2007). These structures were later mapped onto a reference brain generated by averaging 1231 individual STP datasets (Oh et al., 2014). The atlas is still under active development and both segmentations and raw image data are freely available in digital form. It has hence become the basis for many studies requiring segmentation of 3D fluorescence microscopy data (Kim et al., 2014b; Vousden et al., 2015).

1.4.3 Standardising Automated Segmentation

While several studies have used automated image registration on whole-brain 3D fluorescence data, the quality of the resulting segmentations has not yet been adequately quantified. Furthermore there have been no attempts to establish a standardised image registration pipeline for 3D fluorescence data. Past publications either omitted details about the software and parameters used for registration (Lein et al., 2007; Oh et al., 2014) or used closed in-house pipelines based on open source registration tools with only partially published parameters (Menegas et al., 2015; Vousden et al., 2015). One aim of this thesis was hence to establish a fast, open and validated pipeline for automated segmentation of 3D fluorescence mouse brain data.

To accomplish this, automated mouse atlas propagation (aMAP) was developed. It provides an interface to NiftyReg (Modat et al., 2010), a fast MRI registration toolkit for affine and b-spline-based free-form registration (see 2.7.3) that was kindly modified

by Marc Modat to run on larger datasets. The Kim et al. (2014b) atlas was used to evaluate the performance of aMAP. The atlas is based on the original Allen atlas (Lein et al., 2007) and was modified to better match the 3D reference brain template. We further smoothed the structures in the Kim et al atlas along the z-axis (dorso-ventral) to reduce the high-frequency noise in the segmentations along this axis (see 2.7.2). Using the parameters of this study, aMAP was capable of segmenting a complete STP dataset in 40 minutes on a Dell T7500 dual-processor workstation.

To validate the suitability of aMAP, a cohort of 22 neuroscientists was split into two groups and each person was asked to segment 10 structures in 3 brains. The two groups were shown data from different brains, so a total of 6 brains were segmented, each by 11 raters. The segmentation quality of aMAP was then compared to that of the manual segmentations performed by human raters.

1.5 Automated Cell Detection

When mapping connectivity using transsynaptic tracers, it is necessary to detect the precise anatomical location of all fluorescently labelled neurons to evaluate the synaptic connectivity. While automated mapping allows to describe the location of any point in the whole-brain dataset in anatomical terms, the fluorescent cells still need to be detected. Classically this is done by manually marking all fluorescently labelled neurons in the dataset ("cell counting"). However, in our whole-brain tracing experiments, the number of labelled cells can be in the tens of thousands, making manual cell counting extremely laborious and time-consuming. It would thus be highly advantageous to automate this task.

1.5.1 2D Cell Counting Approaches

Detection and even classification of cells in 2D images is routinely used in the analysis of cell culture microscopy images and a variety of free and commercial solutions are available to aid in this task (Abbas et al., 2014). While cell detection tools have classically relied on image filtering algorithms to isolate and detect cells (Carpenter et al., 2006; Malpica et al., 1997; Meyer and Beucher, 1990), machine learning approaches using e.g. random forests (Sommer et al., 2011) or support vector machines (Han et al., 2012; Misselwitz et al., 2010) on a set of parameters calculated from the image data

have recently gained popularity. Despite the high quality of cell detection on cell culture data, accurate and reliable analysis of histological or fluorescence tissue data presents a far more challenging task and has remained elusive (Irshad et al., 2014; Madabhushi and Lee, 2016; Meijering, 2012). However, recent results on histological data using deep learning have shown very promising results (Xue et al., 2016).

Despite the fact that whole-brain scanning using STP or LSFM generates 3D datasets, automated 2D cell counting has been employed to approximate the number of cells in whole-brain STP data (Kim et al., 2014b) by multiplying the detected cell number with a correction factor to account for missed cells.

1.5.2 3D Cell Counting Approaches

The widespread adoption of 3D fluorescence imaging methods such as confocal microscopy has led to an increased interest in 3D cell detection. As a result, several 3D cell detection algorithms have been developed for these relatively small high-resolution datasets (LaTorre et al., 2013; Oberlaender et al., 2009; Toyoshima et al., 2016).

Whole-brain imaging, however, presents a new and unique challenge due to the fact that the amount of data is orders of magnitude larger (~2.5TB per brain for our STP scans), while resolution and image quality are lower. For example, the data for this thesis was acquired using a 10x lens, as opposed to the 40x or 62x lenses used for automated cell counting in confocal images (LaTorre et al., 2013; Oberlaender et al., 2009; Toyoshima et al., 2016). As a result, these 3D cell counters cannot be directly applied to STP data.

This has led to the development of 3D cell counting algorithms specially tailored for whole-brain microscopy (Menegas et al., 2015; Renier et al., 2016; Vousden et al., 2015). Unfortunately, two of these algorithms are in-house pipelines without published quantification of accuracy (Menegas et al., 2015; Vousden et al., 2015). While the recently published cell counting tool by Renier et al. (2016) is freely available, it has only been used on relatively small low-resolution datasets (voxel size of 4µm, 20GB/brain), and its performance has only been quantified in a small region of one dataset (203x203x65 voxels).

Thus, there is a need for a validated open source 3D cell counting method aimed at large whole-brain datasets. These datasets are of lower resolution than the data typically generated with confocal microscopes, and can contain a number of artefacts of similar

shape and size as a cell (e.g. due to contaminations or background fluorescence). As a result, detailed morphological analysis, e.g. using machine learning is required for accurate classification, however the huge size of the datasets prevents such a computationally costly analysis on a complete whole-brain STP dataset.

1.6 Machine Learning in Image Analysis

Machine learning algorithms are a class of algorithms that – rather than following a static set of instructions with predefined parameters – can modify their internal parameters to improve their performance at a defined task in response to repeated execution of the task. Learning algorithms started in the 1950s and were driven by the goal of artificial intelligence (Feigenbaum and Feldman, 1963). To allow computers to "perceive" their environment, learning algorithms were successfully applied to object recognition tasks such as the recognition of handwritten characters (Uhr and Vossler, 1961). To understand the motivation for using machine learning instead of a fully deterministic algorithm, consider the task of recognising a boat in a picture. A boat could be a catamaran, canoe, rubber boat or any other type of boat. It could be on water (partially submerged), on the beach or mounted on a trailer. In addition the colour, illumination profile and size of the boat will vary from picture to picture. This large variability makes it extremely challenging to develop an accurate deterministic algorithm for the task. However, given enough training data, an adequately designed learning algorithm should be able to determine a matching set of parameters to detect the object. Learning such a classification task is known as supervised learning, as the algorithm is given a set of images with a corresponding set of labels that describes its content, and the goal is to accurately predict the label, given the data[2]. The algorithm learns by adapting its parameters in response to the errors during training.

1.6.1 Parameter-Based Machine Learning

Most early machine learning algorithms that achieved success in object recognition tasks did not process raw image data. Instead they relied on human experts to define and calculate a set of features from the images that sufficiently describe their content (Feigenbaum and Feldman, 1963). The machine learning algorithms operated on these

[2]As opposed to unsupervised learning, which is used to find patterns in unlabeled data.

features to try and classify the image. For example, a support vector machine (Hearst et al., 1998) attempts to fit a hyperplane through the feature space, whereas a decision tree will follow a path of binary decisions until it reaches a conclusion about the object in the image (Quinlan, 1986).

While these algorithms modify their parameters during learning to best fit a model to the features, they still essentially rely on external parameters defined by human experts with prior knowledge of the problem to describe the image for them. This has led to the development of a range of specialised image filters and feature detectors that attempt to detect key points while being invariant to e.g. scale or differences in exposure (Bay et al., 2008; Lowe, 2004; Zhang et al., 2007). It would be desirable, however, to have a machine learning algorithm that is able to directly recognize features in and learn from the raw data, removing the need to manually develop mathematical feature descriptors.

1.6.2 Artificial Neuronal Networks

As machine learning was first introduced to try to create artificial intelligence, it is not surprising that artificial neuronal networks (ANN) were amongst the first machine learning algorithms to be used (Farley and Clark, 1954; Rochester et al., 1956). Arguably the most notable early neuronal model that is still in use today (in an extended form) is the perceptron (Rosenblatt, 1958). Its neuronal model states that given a real valued *input vector* x, a *weights vector* w and a *bias* b, the model will output 1 if $w \cdot x + b > 0$ and 0 otherwise. The perceptron was eventually developed in hardware, with a single layer of 512 artificial neurons (AN), randomly connected to 20x20 photocells (Bishop, 2006). The machine was trained to perform object detection in images projected onto the photocells and "learning" was implemented via motorised potentiometers that encoded the weights matrix w.

However, single layer perceptrons were later proven to only be capable of learning patterns that were separable by simple linear equations (Minsky and Papert, 1969), leading to a decline in interest in the techniques. While it was known that an ANN with multiple stacked layers of ANs ("hidden layers", as they are "hidden" between the input and output) could learn more complex patterns (Minsky and Papert, 1969), such networks required large computational power, which was not available at the time.

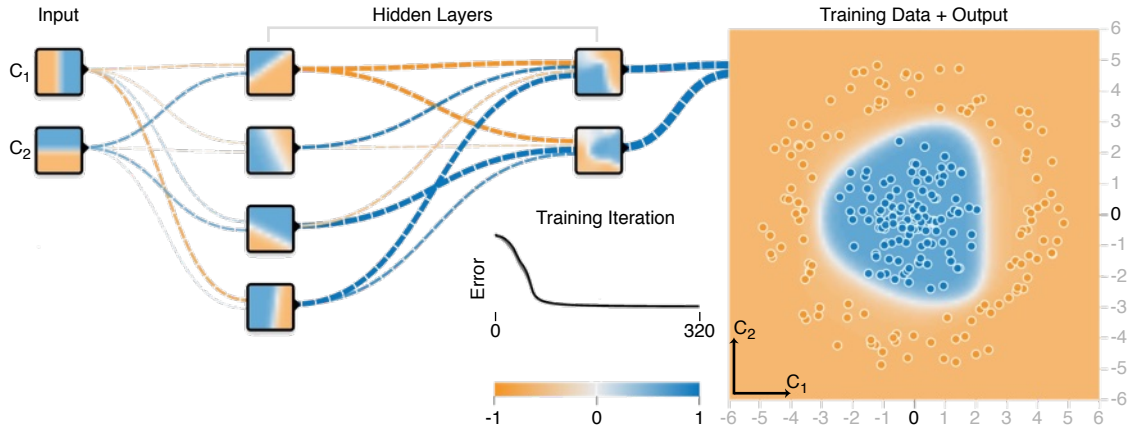


Figure 1-3: Illustration of a simple ANN

A diagram of an ANN with 2 hidden layers using a hyperbolic tangent activation function, performing classification (blue/orange) of a 2-dimensional input value. The dots in the graph show the training data, with the color representing their class. Each box represents an AN; the colour within the box illustrates an AN's output mapped to its input space. The lines between individual ANs represent connections, with the colour showing their sign (blue: positive, orange: negative) and the thickness illustrating the weight of the connection. Each AN calculates the weighted sum of all its inputs, with its output being the hyperbolic tangent of the summed input. The colour map in the output plot shows the segmentation of the data space by the neuronal network. The error plot shows the error as a function of the training iteration. Visualisation was created using the tensorflow playground (playground.tensorflow.org).

1.6.3 Deep Learning

ANNs have regained popularity from the 1980s onwards due to continuing increases in processing power and algorithmic optimisations. In particular the introduction of backpropagation, a method whereby the output error is proportionally applied to the weights of each AN (Werbos, 1974), has been crucial, as it allowed to train neural networks by gradient descent, resulting in vastly improved trainability (Schmidhuber, 2015). More recently, the introduction of libraries such as CUDA (Nickolls et al., 2008) and OpenCL (Stone et al., 2010) that allow to use Graphics cards (GPUs) as affordable parallel processing units allowed to develop and train extremely complex and deep ANNs with tens to hundreds of hidden layers on regular consumer hardware rather than large and expensive cluster systems (Krizhevsky et al., 2012).

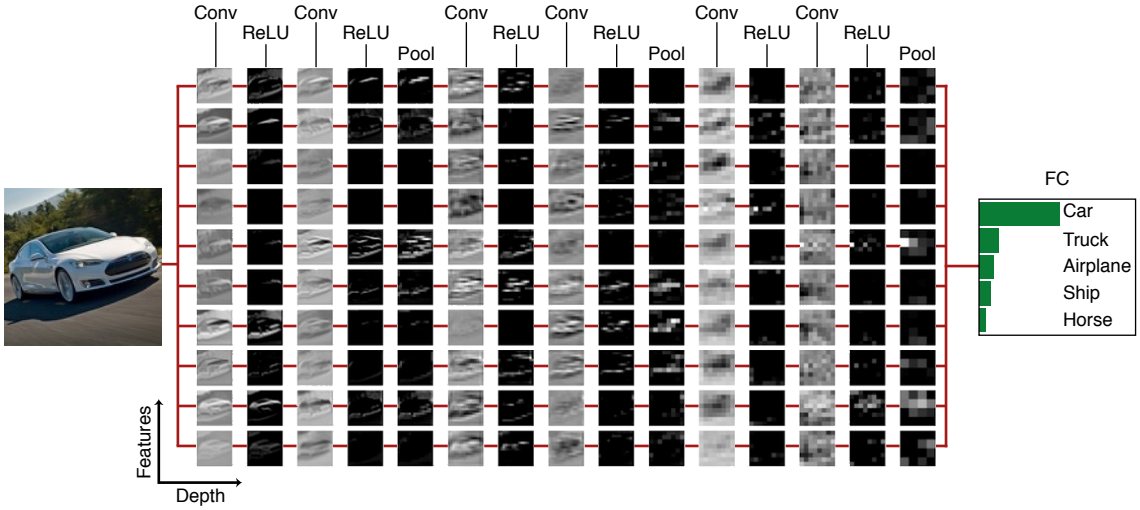


Figure 1-4: Illustration of a convolutional network

Convolutional network with 6 hidden layers (Conv, horizontal axis) and 10 features (vertical axis). The small images show the activation strength for each AN in each layer, with each pixel representing an AN. Each AN in a Conv layer receives input from all 10 ANs at the same location in the previous layer. In this example, the AN is split to show both the result of the weighted summation (Conv) and non-linear activation (rectified linear unit, ReLU, Nair and Hinton (2010)). Spatial reduction is achieved via maximum binning (Pool), with a bin size of 2x2. A grayscale version of the input image is presented to the receptive fields of all 10 feature layers. Conv: convolutional layer; ReLU: rectified linear unit; Pool: max pooling layer; FC: fully connected layer. Adapted from the Stanford University CS231n course material (cs231n.github.io).

Working with these multi-layered (deep) networks has coined the term deep learning[3]. Deep ANNs are currently dominating the field of visual object recognition, having won every ImageNet challenge since 2012 (Russakovsky et al., 2015). Therefore, they are ideal candidates for a cell detection algorithm.

[3] There are also other deep (multilayered) learning architectures that do not rely on artificial neurons, such as multilayer kernel machines (Cho, Y. and Saul, L.K. 2009), however they are not as widely used.

1.6.3.1 Basic Components of ANNs

The first layer in an ANN is the input layer, which represents the data. In the case of 3D image classification, each point in the input layer represents an individual voxel. If the images are sufficiently small (e.g. 50x50x50), the input layer can cover the whole image. If the images are large, the input layer usually represents a sliding window that is moved across the dataset.

The arguably most important layers are the hidden layers, artificial neuronal layers that lie between the input and output layer. Figure 1-3 shows an example of a simple ANN with a single 2-dimensional input (C_1/C_2) and two hidden layers. The training data in this example is artificial, but one could for example imagine that the input is a single pixel (normalised to a mean of 0) of a two-channel fluorescence microscopy image of tissue stained with two different antibodies. The task of the net would be to classify whether the two colour intensities (C_1 and C_2) are similar enough to represent a co-staining at this location. Each point in the training data plot represents a labelled input, with blue denoting co-localisation and orange no co-localisation. The ANs of the first hidden layer receive the C_1/C_2 values as input and – like the neurons in the original perceptron – calculate a weighted sum of the input. The output of the neuron is then calculated from that value using a non-linear activation function. In the case of the original perceptron, the activation function was a simple thresholded binary all-or-nothing response while this example uses the hyperbolic tangent. The non-linearity in the activation function is crucial, as an ANN would otherwise simply represent a linear model (a composition of linear functions always results in a linear function).

The ANs of the second hidden layer now receive the output of the first hidden layer and likewise perform weighted summation and activation. The output of the second hidden layer is then summed to represent the final output of the network. The individual activity patterns of the ANs in Figure 1-3 illustrate nicely how complex patterns arise using multiple layers of simple summation followed by nonlinear activation.

In the example in Figure 1-3, a single two-dimensional input (e.g. a 2-colour pixel) is routed to four ANs in the first hidden layer, with each of the four ANs representing an abstract feature that is used to classify that point. ANNs for image classification use an extension of this method called convolutional layer (Figure 1-4, Le Cun et al. (1990)), where each pixel of the image is routed to a number of ANs, with each neuron

representing an abstract feature of that particular pixel (Figure 1-4, Conv layers, vertical axis). Alternatively, an AN can receive the input of a group of neighbouring pixels, enabling the detection of spatial patterns, which can be seen as an analogue to a receptive field in the visual system.

Many convolutional ANNs gradually reduce the spatial size of the data, while increasing the number of features (Krizhevsky et al., 2012; Szegedy et al., 2015; Zeiler et al., 2011). Spatial reduction can be achieved either by using receptive fields with the distance of their centres (stride) larger than 1 or by using pooling layers, which use a simple binning operation (e.g. mean or max) to reduce the size of the data (Figure 1-4, Pool layers).

In a classification task, the information from the final convolutional layer is projected onto one (or multiple) fully connected (FC) layers. Here, every AN receives the input of all ANs of the previous layer, irrespective of their spatial position, to combine all information and determine the class of the image (Figure 1-4, FC layer). The final FC layer contains one AN per class, and the segmentation result is defined by the AN with the strongest activation (Figure 1-4, FC shows the relative activation strength of the 5 strongest ANs).

1.6.3.2 Training ANNs

ANNs can either be trained in a supervised or unsupervised manner. Unsupervised learning uses unlabelled data and is generally used to find consistent patterns or structures in the input data. While unsupervised learning can be used to pre-train an ANN for image classification (Bengio et al., 2007), this method has not been used in the thesis.

During supervised learning, the ANN is presented with an image and its associated class label, which defines the optimal result of the last FC layer. The ANN then classifies the image and calculates the difference between the classification result and the optimal result. This serves as input to a loss function (also called cost or error function), which is then “backpropagated” through the network by calculating the partial derivatives with respect to the weights for each neuron in each layer. This results in a measure of error for each weight of each neuron, which is used to update the weights for the next image (gradient descent). In regular intervals, the network is evaluated by classifying images from a validation dataset without backpropagating the error or updating the weights.

Training is usually halted when the error on the validation dataset increases (Sarle, 1996).

To prevent the network from oscillating, images can be presented in batches and the weights are only updated after a batch of images. Furthermore, a learning rate <1 is multiplied with the errors to reduce their influence. Finally, the learning rate is also often decreased with increasing training time, which results in gradually smaller changes to the weights to improve the stability of the network (Bottou, 2012). While training is a computationally expensive task that, depending on the complexity of the problem, can take days or weeks even on high performance computers (Iandola et al., 2015), applying a trained network to input data is relatively fast. The network from Figure 1-4 for example was implemented in JavaScript and can be run in a web-browser (http://cs231n.stanford.edu/).

1.7 Applications for Automated Connectivity Analysis

Whole-brain connectivity analysis driven by advances in viral tracing and automated microscopy now makes mapping the whole mouse brain a feasible goal that is actively being pursued (Bohland et al., 2009; Osten and Margrie, 2013). It has already led to a remarkable online resource by the Allen Brain Institute that shows connectivity in the mouse brain mapped using Cre-dependent rAAV in combination with STP (Oh et al., 2014). Transsynaptic tracing methods take the approach one step further by reporting synaptic connectivity. When combined with a potential broader adoption of whole-brain imaging techniques and automated data analyses they have the potential to rapidly advance our knowledge about connectivity and function of the brain.

1.7.1 Cell-Type Specific Mapping

Advances in the development of transgenic mouse lines have led to the availability of a large variety of so-called Cre lines, transgenic mice that express Cre recombinase (Sauer and Henderson, 1988) in genetically defined populations of neurons (Josh Huang and Zeng, 2013; Orban et al., 1992). Using RV and Cre-dependent rAAV helper viruses in these mouse lines allows targeting cells not only by location but also by cell type (Josh Huang and Zeng, 2013). This can be used to distinguish and unravel distinct neuronal circuits in the same anatomical area, as shown by Vélez-Fort et al. (2014),

who used Cre-dependent RV tracing in combination with morphological reconstruction and electrophysiological characterisation to describe two functionally distinct microcircuits within layer 6 of the primary visual cortex.

1.7.2 Mouse Models of Disease

Animal models have become a crucial tool for understanding human illnesses by enabling the study of severe conditions on a molecular and systemic level. The unique ability to mimic, investigate and test treatments for diseases in a way that would be impossible with human subjects has led to breakthroughs in many fields, from cancer (Semenza, 2003) to HIV (Klein et al., 2013) and spinal cord injuries (Wenger et al., 2014). Despite that, our understanding of and treatment options for many neuropathological disorders remain limited (Nestler and Hyman, 2010).

Advances in genetic screening have led to the discovery of potential targets for many disorders affecting the nervous system and have thus enabled the development of genetically engineered mouse models for a variety of such disorders including Parkinson's (Przedborski and Vila, 2003), Alzheimer's (Götz and Ittner, 2008), Down's (Li et al., 2007), Depression (Kalueff et al., 2007), Schizophrenia (Belforte et al., 2010) and Autism (Peça et al., 2011). However, the ability to evaluate a complex behavioural phenotype such as depression or schizophrenia in a mouse model is limited at best (Nestler and Hyman, 2010), making evaluation of the validity of these disease models difficult. This is compounded by the fact that neuropsychiatric conditions in particular can present themselves in a very heterogeneous way, which can lead to two individuals being diagnosed with the same disorder despite having different symptoms with little overlap (Nestler and Hyman, 2010).

As these diseases are known to have an impact on brain wiring, for example with altered spine number and morphology in Down's syndrome (Contestabile et al., 2010) and abnormalities in the wiring of the cerebral cortex in Autism (Geschwind and Levitt, 2007), connectivity analysis in these models presents a unique opportunity to shed further light on the impact that these diseases have on the level of individual neuronal circuits. While low-resolution rfMRI data from human subjects is readily available (Broyd et al., 2009), functional connectivity, as measured by rfMRI does not equal anatomic connectivity. It does, however provide a set of possible targets for detailed

connectivity analysis at cellular resolution to complement the rfMRI data and further improve our understanding of the underlying network changes in neuropathological disorders.

1.7.3 Reliability of Connectivity

When discussing connectivity, many experiments are designed on the assumption that connectivity is stereotypical and similar between individuals. Even the large-scale connectivity mapping project by Kim et al. (2014b) rarely includes more than a single injection with the same parameters (mouse line, injection coordinate, volume and batch). However, the assumption that all mice (or men) are essentially the same may be overly simplistic. The high-throughput and precise quantification capabilities of automated mapping would allow investigation of subject-to-subject variability in neuronal connectivity.

1.8 Aim of the Thesis

The aim of the thesis was to develop and validate an automated pipeline for tracing whole-brain connectivity of defined populations of neurons in 3D STP datasets of the mouse brain. To achieve this goal, two essential pieces of software were developed and evaluated for the following tasks:

1. Automated image registration and segmentation
2. Automated cell counting

Chapter 2. Materials & Methods

2.1 Mouse Lines

All mice were on a C57BL/6 background. The transgenic Cre-reporter mice used in this study were of type Ntsr1Cre to target layer 6 cortico-thalamic projection neurons in the primary visual cortex (Vélez-Fort et al., 2014) and GAD2-IRES-Cre to label GAD2 positive interneurons in the primary visual cortex (Harris et al., 2015).

2.2 Viral Vectors

The rabies virus used in this study was an EnvA-pseudotyped SAD19 strain rabies expressing mCherry instead of the rabies glycoprotein (EnvA-RV-mCherry). It was produced according to previously published methods (Vélez-Fort et al., 2014).
The EnvA-RV-mCherry was trans-complemented by prior stereotaxic injection of a mixture of two recombinant adeno-associated virus vectors, serotype 8 (rAAV8). The first rAAV8 expressed an E2A-linked fusion protein of Cerulean and the SAD19 rabies glycoprotein (Cerulean-E2A-SADB19RG) under the control of a Synapsin promoter and a Flex site (based on Addgene #49101), while the second rAAV8 expressed EGFP-E2A-TVA under the control of the ef1a promoter and a Flex site (Addgene #26198). For injection, the two rAAV8 were mixed in a 2:1 ratio (RG to TVA).

2.3 Stereotaxic Injections

All procedures were carried out in accordance with UK Home Office regulations (Animal Welfare Act 2006) and the local animal ethics committee. Briefly, animals were anaesthetised using intraperitoneal injection of Ketamine/Xylazine (2:1 mixture) and a minimal craniotomy was performed over the target area. Injections were carried using long-shanked volume-calibrated pipettes (Blaubrand, Brand GmbH, Germany) that were pulled and broken to a tip diameter of ~7µm. Injection pipettes were tip-filled under negative pressure and injected using positive pressure, ensuring an injection duration of ~3min. V1 injections were carried out on an in-vivo patch-clamp setup

calibrated to target V1. Craniotomies were sealed post injection using Kwik-Cast Sealant (World Precision Instruments). Mice were injected with ~20nl of rAAV followed 3 days later by ~50nl of RV. Animals were sacrificed 10 days after RV injection.

2.4 Tissue Preparation

Mice were perfused trans-cardially with cold 4% PFA-solution under deep general anaesthesia (Ketamine/Xylazine 2:1). Brains were then removed, post-fixed in 4% PFA for at least 24h and embedded in 4% type 1 agarose (Sigma-Aldrich).

2.5 Data Collection

Images were acquired in the coronal plane using a TissueCyte 1000 STP (TissueVision, Cambridge MA, USA) with an XlumplanFl 10x water immersion objective (Olympus Corporation, Tokyo, Japan) and a Mai Tai DeepSee 2-photon excitation laser (Spectra-Physics, Santa Clara, USA) at an excitation wavelength of 800nm. Images were acquired at 16bit with a resolution of 1664x1664 (1µm/pixel, 0.4µs/pixel) per single image tile with each optical layer consisting of 9x6 image tiles. After acquiring 10 optical layers with a z-step of 5µm, a 50µm section was cut automatically. This process was repeated until the complete brain was imaged, resulting in 300-320 tissue sections per brain.

2.6 Data Handling

Acquired image tiles were processed with an automated image processing pipeline (tvPy) written in Python. Individual tiles were cropped to a resolution of 1568x1568, rotated 90˚ clockwise and corrected for uneven illumination by generating a correction tile that is the average of all image tiles of the corresponding optical layer in each of the ~300 physical slices. The resulting 10 correction tiles (1 per optical layer) were then subtracted from the individual image tiles of the corresponding optical layers. The processed image tiles were stitched into planes using a custom version of the stitching plugin originally developed by Preibisch et al. (2009), which was modified to stitch the images based on the XY-stage coordinate output of the TissueCyte STP.

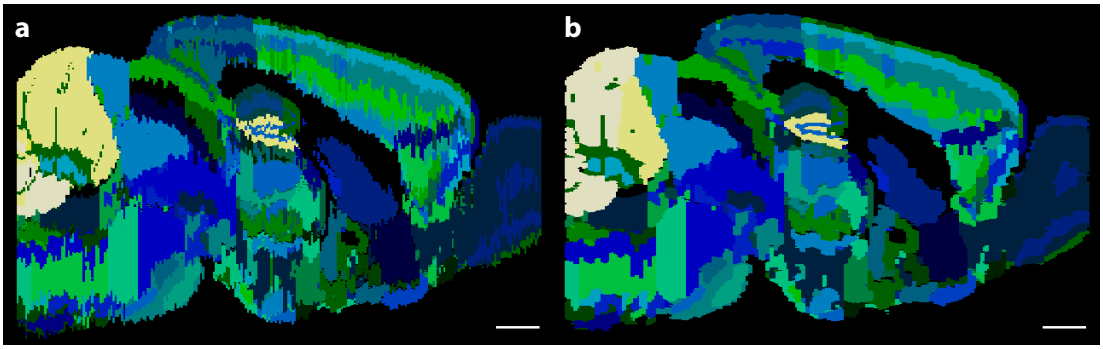


Figure 2-1: High frequency noise along the z-axis of the segmentation dataset

a) Image of a sagittal section through the segmentation of the Kim et al. atlas (scaled to isotropic size), with different anatomical areas displayed in a false color map. High frequency noise is visible in the structure boundaries from anterior to posterior caused by artifacts stemming from the fact that the atlas is based on 2D segmentations of individual coronal cryostat sections. b) The same image after two iterations of Gaussian smoothing of the structures (radius 0.5 voxels at a voxel size of 20x20x50µm), showing noticeably reduced noise. Scale bars = 1mm.

2.7 Automated Segmentation using aMAP

2.7.1 Data Preparation

For automated segmentation, each STP scan was first smoothed along the z-axis (Gaussian, s.d. of 5 voxels) to reduce the influence of depth-dependent illumination changes and then downsampled to a voxel size of 12.5µm isotropic. To enable registration with NiftyReg, the STP data was converted to the Nifti file format using MATLAB (Mathworks) with the "Tools for NIftI and ANALYZE image" (http://uk.mathworks.com/matlabcentral/fileexchange/8797-tools-for-nifti-and-analyze-image). Both manual and automated segmentation were carried out either on background fluorescence (n=5 brains) or the RFP signal of a sparsely labelled animal (n=1 brain).

2.7.2 Atlas Preparation

For use with aMAP, the segmentations from the atlas developed by Kim et al. (2014b) were smoothed twice using a Gaussian kernel (s.d. of 0.5 voxels). This was done to reduce the impact of high-frequency noise along the z-axis (Figure 2-1).

2.7.3 Registration Using NiftyReg

Image registration with NiftyReg, which forms the basis of aMAP, is a two-step process consisting of an affine image registration followed by a free-form registration. During this process, the STP data remains unchanged while the average brain is modified to maximise the similarity to the STP data. Both registration steps use a pyramidal approach, meaning the data is downsampled multiple times by a factor of 2. The registration is then performed on the lowest resolution before moving up to the next larger version to prioritise global matching of structures and avoid local minima.

2.7.3.1 Affine Registration Using reg_aladin

Affine registration was used to obtain a rough overall fit between the atlas and our STP data. The atlas data was modified using the set of affine transformations (translation, rotation, scaling, shearing) to obtain maximum similarity between the two datasets, where similarity was calculated using an iterative symmetric block-matching approach. Here, both the atlas and STP data were divided into small blocks and the algorithm attempts to find matching blocks in the two datasets using normalised cross correlation. The distances between all matching blocks were calculated and the affine transformation that minimises these distances was obtained using least trimmed squares regression (Modat et al., 2014).

2.7.3.2 Free-Form Registration Using reg_f3d

Free-form registration was used to optimise the fit after affine registration. Here, a regular grid of control points is placed over the image. These control points were moved during the registration and influence the area around them via a b-spline relationship. The goal or the algorithm is to maximise the registration score, which is a combined measure of image similarity and a regularisation term that prevents overfitting. The following parameters of the free-form image registration are known to have a large impact on the result and were optimised on 2 out of the 6 STP datasets:

Spacing of the control points of the b-spline grid

Sets the distance between the individual control points of the b-spline grid. A larger spacing enforces more global transformations while a smaller spacing allows more local registration. Set to 10 Voxels.

Similarity function

The first term influencing the registration score during optimisation is the similarity between the target image (STP data) and the "floating" image (reference brain of the atlas), which is calculated for each iteration of the optimisation. The similarity function defines how similarity is calculated and can be either set to normalised mutual information (NMI, with adjustable number of bins) or locally normalised cross-correlation (with adjustable Gaussian kernel size for normalisation). Set to NMI with 128 bins.

Bending Energy Weight

The second term influencing the registration score is the second derivative of the grid point translation, evaluated at each point of the b-spline grid ("bending energy"). It acts as a regularization term and penalises high frequency movement of the b-spline grid points. The bending energy weight (BE) shifts the relative weight of these two terms as follows:

$$registrationScore = (1 - BE) * similarity - BE * bendingEnergy$$

Set to 0.95.

Total number of steps in the downsampling pyramid

This parameter controls how many downsampling steps are generated for the pyramidal registration approach (see 2.7.3). Set to 6.

Number of computed steps in the downsampling pyramid

As discussed above, registration starts on the smallest dataset of the downsampling pyramid. This parameter controls how many steps are used to compute the final registration. It is, for example, possible to increase registration speed by calculating 5 downsampling steps but only register 4 of them (omitting the registration on the full-sized data). Set to 4.

The 2 datasets used for optimisation were excluded from all analyses of aMAP's performance.

2.8 Manual Segmentations

22 neuroscientists from the former Division of Neurophysiology at the National Institute of Medical Research were recruited as human raters for this task. These included P.I.s, postdocs, PhD students and technicians. Raters were asked to use the online version of the Allen mouse brain atlas to segment 10 different brain structures. Each rater was asked to segment each structure on STP datasets from 3 different animals showing background fluorescence. In addition, unbeknownst to the raters, all images from one animal were presented again to assess the reliability of segmentation, resulting in a total of 40 segmentation tasks per rater. First and repeated presentation of the same data were spaced at least 20 segmentation tasks apart. The 22 raters were split into two groups that were presented data from different animals, leading to a total of 6 brains being segmented by 11 raters each.

2.8.1 Data Preparation

The STP data was rigidly aligned to the average brain of the Allen Mouse Brain atlas, to ensure correct alignment in the coronal plane. The transformation matrices for this alignment were determined using reg_aladin on STP data prepared as described above and then applied to the full-resolution STP images using MATLAB (MathWorks)

2.8.2 Choice of Brain Structures

Brain structures were chosen to cover a large range of areas across the dorsoventral and anterior-posterior axes of the brain and varying levels of anticipated segmentation difficulty. The brain structures included in the analysis were Anterior Cingulate Area (ACA); Anterior Hypothalamic Nucleus (AHN); Dentate Gyrus, granule cell layer (DG-sg) [4] ; Medial Vestibular Nucleus (MV); Retrosplenial Cortex (RSP); Primary Somatosensory Cortex (SSp); Subiculum (SUB); Primary Visual Cortex (VISp); Secondary Visual Cortex, anteriomedial part (VISam) and Ventral Posteromedial Nucleus of the Thalamus (VPM).

[4] Analysis of the data indicated strong influence of the z-choice on the human DG-sg segmentations. This structure was therefore excluded from segmentation analysis (see Results)

2.8.3 Data Presentation

STP images were presented on a Wacom Cintiq screen to allow segmentation with a digital pen. The coronal Allen brain atlas plates were presented on a separate computer and raters were free to browse them as required. For each brain structure, the raters were presented with a coronal z-stack consisting of 40 STP images, acquired 15µm apart. Raters were then given an atlas plate number and structure name from the Allen brain atlas, asked to find the image of the STP stack that best corresponded to the atlas plate (z-choice) and segment the given structure on that image. Data were presented to each of the two groups in four blocks of 10 structures. Data presentation and storage of the raters' segmentations was handled by a custom ImageJ plugin.

2.8.4 Post-Processing of Manual Segmentations

Segmentation outlines were manually cleaned by removing artefacts such as isolated small areas that occur when a rater accidentally touches an unrelated part of the image with the digitizer pen during draw mode. 5 of 880 segmentations were performed on the wrong structure or hemisphere and were thus discarded. The remaining segmentation outlines were downscaled to an x-y pixel size of 4µm and converted to filled binary images.

2.9 Assessment of Segmentation Performance

2.9.1 Euclidean Landmark Distance

The following 10 anatomical landmarks, as used in the study by Kim et al. (2014b) and defined by the Waxholm Space (Johnson et al., 2010), were marked in the average brain of the Kim et al. atlas and the downscaled version of each STP dataset before registration: frontal middle 1; frontal right 2; frontal left 2; anterior commissure right; anterior commissure left; corpus callosum middle; hippocampus middle; interpeduncular nucleus right; interpeduncular nucleus middle; interpeduncular nucleus left. In one STP dataset, hippocampus middle was omitted due to an imaging artefact in that area. The free-form registration was then calculated for each of the 6 brains using 18 different BE values in a range from 0.2 to 0.99.

2.9.2 Scoring Using Consensus Segmentations

Since there is no 'ground truth' segmentation that could be used to assess the quality of individual segmentations, all segmentations for a given target structure were compared to a 'consensus segmentation' derived from all manual segmentations of that structure using STAPLE (Warfield et al., 2004). STAPLE is an iterative algorithm and aims to simultaneously assess the 'quality' of each segmentation and the quality-weighted consensus segmentation of all segmentations. The quality of a segmentation is derived from its overlap with the consensus segmentation and is initialised to equal levels for all segmentations. After the first iteration, segmentations with low overlap receive a penalty in their quality rating while segmentations with high overlap are assigned a higher quality value. This process is then repeated until convergence. STAPLE consensus structures were generated from manual segmentations using NiftySeg (Jorge Cardoso et al., 2013). Both manual and automated segmentations were scored against the consensus segmentation using the Dice score (Dice, 1945). As supplementary measures, consensus segmentations were also generated using shape-based averaging (SBA, Rohlfing and Maurer (2007)), which calculates the geometric mean of multiple areas. All segmentations were also scored using the Hausdorff distance, which is defined as the longest distance between any point on one set and its closest neighbour in the other set. It measures the maximum distance between two segmentations and is hence a good indicator of segmentation artefacts. All segmentations were scored using a custom analysis pipeline written in Matlab.

2.9.3 Comparison of 2D Manual Segmentations and 3D aMAP Segmentations

While aMAP generates 3D segmentations of the complete brain, manual segmentations of complete anatomical structures in 3D are not achievable in a reasonable amount of time. To score aMAP's performance, its 3D segmentations were cut coronally to generate comparable 2D aMAP segmentations. The section with the highest score was used for each anatomical structure.

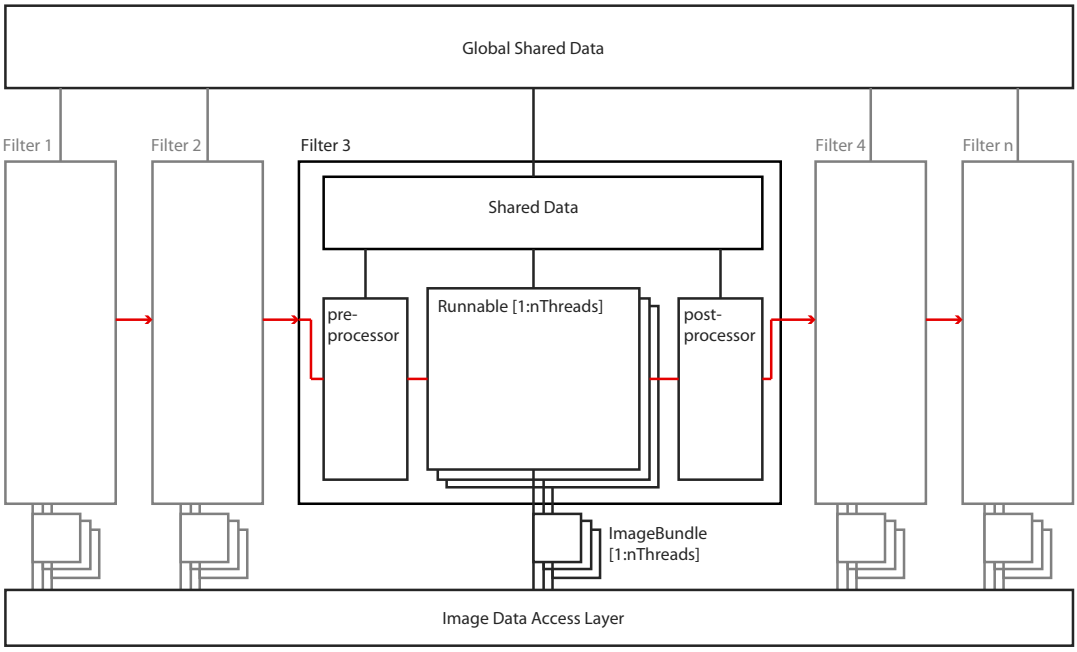


Figure 2-2: Schematic of FACCT filter design

FACCT applies a sequence of **filter**s (red line) to the image data to locate cells in the dataset. A **global shared data** object is used for storing information that is relevant for multiple filters or results needed by a later filter. Each filter is comprised of 4 components: a **pre-processor** for initialisation, multiple **runnables** that process different regions of the dataset in parallel, a **post-processor** that performs clean-up tasks and can store results for later filters and a **shared data** object that contains persistent data and defines the required user input. Images are provided to each runnable by the **data access layer** via an **ImageBundle** that provides read/write-access to the relevant image data of each runnable.

2.10 Manual Cell Counting

Cells were counted manually on sub-volumes of the STP data using either a custom modified version of the ImageJ (Schneider et al., 2012) cell counter plugin originally developed by Kurt De Vos, or MASIV, a Matlab visualisation interface for large 3D image datasets developed in our lab by Alex Brown.

2.11 Automated Neuron Detection using FACCT

FACCT consists of a modular toolkit written in Java and an ANN implemented in Caffe. The java toolkit enables simple development of "filters", which are modules that

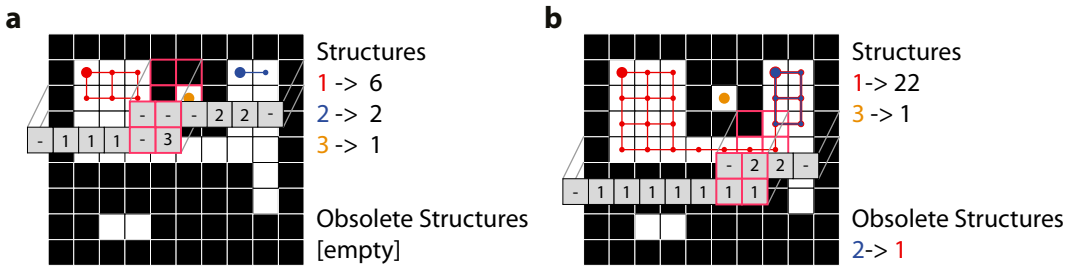


Figure 2-3: Ring buffer and connected structure analysis

Illustration of two stages in the analysis of connected structures in a hypothetical tile. The red 2x2 grid denotes the analysis kernel, the grey overlay shows the ring buffer that stores the structure IDs that may still be relevant for analysis. a) At this point, the algorithm has detected 3 different structures (red, orange, blue). Their sizes are stored in a map (Structures) along with their ID. b) At this point, structures 1 and 2 (red and blue) are detected to be connected. The size of structure 2 is added to structure 1 and structure 2 is marked as obsolete and deleted from the Structures map.

process a series of images in a linear fashion and integrates with ImageJ, enabling access to a large range of well-established image processing routines. Image data is processed using sequence of multiple filters to locate potential cells in the dataset which are than analysed by the ANN. Data access code is completely separated from the filters, allowing to process images in memory, stored layer-by-layer on hard disk or even data from specialised multi-folder arrangements without modification of the actual image processing code.

2.11.1 FACCT Filter Design

The filter is the core element of FACCT (Figure 2-2). The purpose of each FACCT filter is to further reduce the volume of "data of interest" until only marked cells remain. A filter can define user-inputs (e.g. a positive integer for cell size, or a directory for logging), which are requested from the user during interactive operation or read from the script command during scripted runs.

2.11.2 Parallel Processing Model

Parallel processing of image data is realised via the "runnable" of each filter: The STP data is split along the z-axis into as many parts as there are processors available and a runnable object is generated for each of these sub-regions. To prevent resource conflicts, runnables run independently on different parts of the STP data and their results are combined during the post-processing step.

2.11.3 Tiled Processing System

To cope with uneven illumination, the initial two filters (tile classifier and thresholder) split the images into small 2D tiles that are analysed independently. Each tile has twice the height and twice the width of the expected large cell diameter, and tiles can be set to partially overlap. The expected large cell diameter has been set to 15µm for this study.

2.11.4 2D/3D Ring Buffers

Many of FACCTs filters operate using a 2D or 3D kernel and require a temporary data structure that is separate from, but aligned with the image data they operate on (e.g. to store the IDs of structures in the image). To ensure high speed and minimal memory usage, these temporary data are held in ring buffers that extend over the area of the image relevant to the filter's kernel (Figure 2-3).

2.11.5 Analysis of Connected Structures

The Tile classifier, thresholder and 3D structure counter all internally depend on the detection of connected binary structures. In all cases detection is carried out by a single pass of a 2x2 (or 2x2x2 in 3D) kernel through the image data. If a new structure is encountered, it is assigned a unique ID and stored in a Map/Dictionary data structure. The structure is then extended or joined to another structure as required (Figure 2-3).

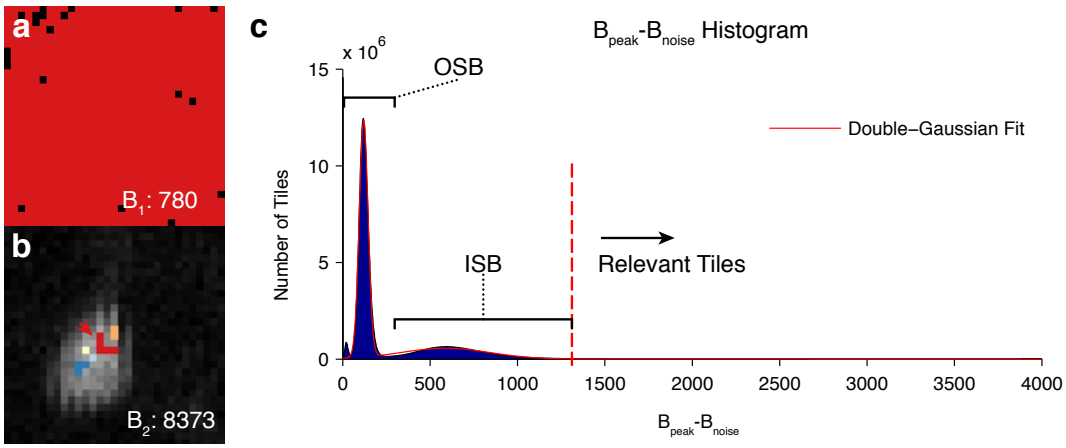


Figure 2-4: Tile classifier schematic

a) An exemplary tile from an STP dataset with the threshold set to B_{noise}, the brightest value which is darker than 98% of all pixels. Thresholded pixels are shown in red. b) The same tile with the threshold set to B_{peak}, the darkest value that generates a 4-pixel structure (red arrow). c) Histogram of B_{np} (the difference between B_{peak} and B_{noise}), computed from all tiles of a complete STP dataset. A double Gaussian fit is marked in red. Only values up to 4000 are shown, the maximum B_{np} was 27111. The red dashed line marks the split between tiles that were discarded and tiles that were used for analysis. OSB: Tiles outside the brain without significant amounts of signal; ISB Tiles inside the brain without significant amount of signal.

2.11.6 Cell Counter Modules:

2.11.6.1 Initial Data Reduction Using Fast Tile Analysis

The aim of this filter is to quickly detect and discard areas of the image that contain only background fluorescence. It is a 2D tile-based filter and measures the size of the largest connected structure at different threshold levels for each tile. It uses a divide-and-conquer strategy to find the brightness level B_{noise} of the noise floor (defined as the brightness where the largest structure encompasses at least 98% of the tile, Figure 2-4 a) and the peak brightness B_{peak} (brightness level at which there is a connected structure with a size of at least 4 pixels, Figure 2-4 b). The difference $B_{np} = B_{peak} - B_{noise}$ serves as an indicator of whether a tile is relevant: In a sparsely labelled full brain STP dataset, the histogram of all B_{np} values has a distinct shape with two pronounced peaks (Figure 2-4 c): A narrow, high peak corresponding to the tiles located outside the brain and a

broader peak corresponding to the tiles inside the brain that contain only tissue background. By specifying a minimum $B_{np,}$ which can either be calculated automatically using a double Gaussian fit or set by the user, these background tiles are excluded.

2.11.6.2 Size-Checked Otsu Thresholding

This filter binarises the remaining tiles by individually calculating the optimal threshold for each remaining tile. It is an extension of Otsu's thresholding algorithm (Otsu, 1975) that takes the sizes of the resulting thresholded structures into account. First, Otsu's threshold is calculated on the tile histogram. Otsu's method performs a clustering on the histogram under the assumption that the image is best described by a bimodal distribution. It calculates the threshold that minimises the sum of the intra-class variances for both classes (thresholded and non-thresholded).

In case of low signal levels, this method can result in underestimation of the correct threshold, resulting in structures that are too large to be a cell. To avoid this, the size of the largest structure in the tile is calculated using the same 2D structure analysis module as the previous filter and compared to a predefined area limit (the square of the "expected large cell diameter", see 2.11.3). If the structure is within the size limit for a cell, the threshold is accepted. If the structure is too large, the correct threshold must be in the brightness range above the current threshold and Otsu's method is applied to the subhistogram above the current threshold. This process is repeated until the size of the largest structure in the tile does not exceed the size limit for a cell.

2.11.6.3 Simple Morphological Filter for Noise Suppression

This filter uses the binarised data and fits a 3D spheroid at each position. Each spheroid position that results in an overlap of at least 90% between the thresholded voxels and the spheroid is marked and all other points are removed from further analysis. This results in removal of minor artefacts and structures such as neurites from the data. As each runnable only has access to one image at a time, the filter uses a 3D ring buffer to keep all relevant image data in memory.

2.11.6.4 Structure Counter

The final filter counts and stores all remaining isolated structures from the above steps using a 3D extension of the method described in 2.11.5. Its development was necessary, since the standard 3D structure counter in Fiji (Bolte and Cordelieres, 2006) cannot be used on large datasets. The result of this filter is a list of the centre of mass positions of all cell soma candidates.

2.11.7 Deep Learning Analysis of Structures

To remove false positives from the cell soma candidates, cubes of 50x50x20 voxels are extracted from the original STP data at the cell candidate positions and classified using deep learning. The network used here is based on a ResNet architecture (He et al., 2015a) with 4 individual ResNet blocks connected to a cascade of two fully-connected layers with intermediate normalisation and rectified linear units (ReLU, see Figure 2-5). In a ResNet, the layers are not trained to learn the a desired mapping $F(x)$, but instead learn its residual $R(x) = F(x) - x$. In other words, a layer does not learn to generate new features from the previous layer, but rather how the output of the previous layer needs to be changed to obtain the desired features. This design was motivated by the observation that the addition of further layers to regular convolutional ANNs would at some point lead to a sudden drop in its training performance. This is somewhat counterintuitive, as a deeper network can – by construction – perfectly represent its shallower counterpart if the additional layers perform an identity mapping. A deeper network should thus not perform worse during training unless the solving algorithm is unable to find the optimal solution. He et al. (2015a) hypothesised that residual layers would be easier to optimise by current stochastic gradient descent solvers, because an identity mapping (or a solution close to it) is easy to express with a residual function (the residual of identity equals zero). On the CIFAR-10 image dataset (Krizhevsky, 2009), adding layers to the ResNet resulted in a performance increase for up to 110 layers. A test with 1202 layers resulted in similar training- but worse testing performance, suggesting overfitting. This is in contrast to the non-residual VGG network (Simonyan and Zisserman, 2014), which the ResNet was based on. The non-residual VGG showed a decrease in training performance at 34 layers (He et al., 2015a). Due to the combination of high performance (it has won the 2015 ImageNet object

localisation challenge) and extendibility, the ResNet architecture was chosen for cell detection in FACCT.

The neural network was implemented in a version of Caffe (Jia et al., 2014) that was modified to accept n-dimensional data. All computations were carried out on a Nvidia Titan X GPU (EVGA). For training purposes, cells were defined as locations that had a human count within a distance of 20µm, while false positives were defined as locations that had no human count within a distance of 40µm. To generalise the training data, the training dataset was augmented (Breiman, 1996) by applying any combination of the following operations to the data: 90º rotations, mirroring and brightness modulation by multiplication with a random constant between 0.5 and 1.5.

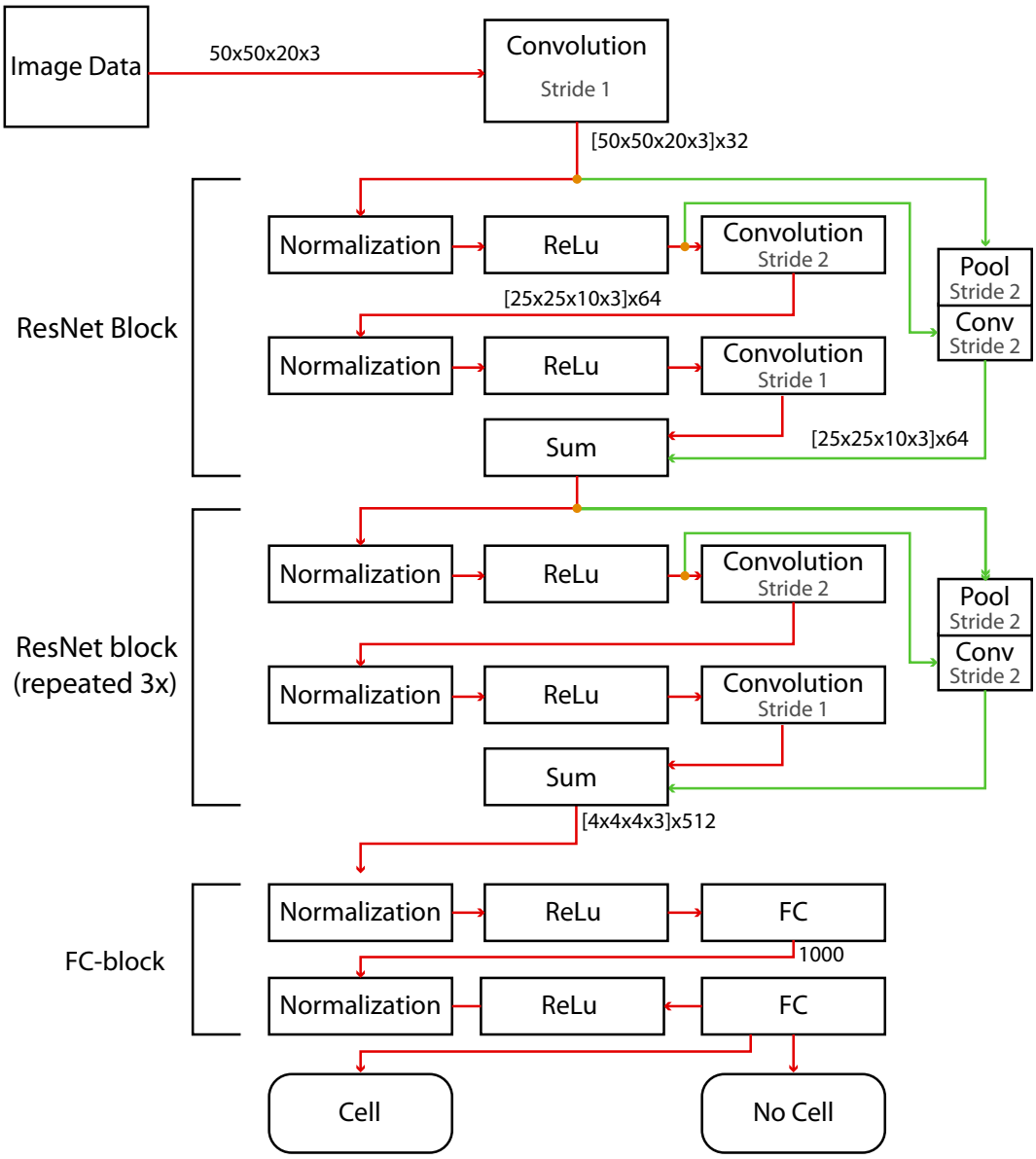


Figure 2-5: ResNet architecture

Simplified diagram of the neural net used in the final cell detection step. The main elements of the network are 4 ResNet blocks that use a shortcut ("identity") connection from the previous layer (green path). The shortcut connection is transformed to the correct output size by an average pooling operation and concatenation with the output of a separate convolutional layer (Conv). This "identity" is then added to the output of the ResNet block to recover the desired result. At the end of the ResNet blocks, two fully connected layers (with normalisation and ReLu) converge the ouput to the two desired values (cell or no cell). Numerical annotations denote the approximate data dimensionality. FC: fully connected layer, ReLu: Rectified Linear Unit.

2.11.8 Evaluation of FACCT Performance

FACCT was evaluated on STP scans of the brains of 6 mice, stereotaxically injected with rAAV/RV into the primary visual cortex (3 Ntsr1-Cre, 3 GAD2-Cre; see 2.1-2.4).

2.11.8.1 Qualitative Analysis

For 3D visualisation of cells detected by FACCT and human raters, the average brain of the Allen common coordinate framework (October 2016 release) was first registered to the individual datasets using aMAP. The positions of cells detected by human raters and FACCT were then displayed on a 3D maximum intensity projection of the registered average brain using Vaa3D (Peng et al., 2014).

2.11.8.2 Count Comparison Using Equal Sampling

To calculate the correlation between the numbers of cells found by human raters and FACCT, the dataset was segmented into overlapping cubes with an edge length of 400µm, with their centres spaced 200µm apart. The number of cells detected by FACCT and the human rater were then counted and plotted for each brain. The regression was calculated using Matlab's fit function with the "bisquare" robust linear fitting algorithm.

2.11.8.3 Count Comparison Using Anatomical Segmentation

The datasets were segmented using aMAP with the Allen common coordinate framework (October 2016 release) and the number of cells detected by FACCT and the human rater per anatomical structure were counted and plotted for each brain. Regression was calculated as above (2.11.8.2).

To analyse the differences in relative cell numbers between FACCT and human raters, the numbers of cells per anatomical region were converted to percentages by dividing the number of cells found in each region by the total number of cells found in that particular brain. The difference in relative cell number between FACCT and the human rater was then calculated for each anatomical region and the results of all brains and regions were plotted grouped by region.

2.11.8.4 Count Comparison on Z-Corrected Data

To evaluate the performance of FACCT on STP data that does not suffer from discontinuities along the anterior-posterior axis, the discontinuities in an area of 1400µm*1200 µm *200 µm including mostly visual cortex were manually corrected. The cells in that area were then marked by 6 human raters. For comparison, cells were marked by FACCT in a version of the same dataset that was extended by 100µm along each axis to allow full extraction of the image cubes needed for deep learning.

As the ResNet used for the previous analyses was mainly trained on cells from STP datasets with strong z-discontinuities, it was retrained on marked cells from a dataset containing relatively few z-discontinuities.

Chapter 3. aMAP: a Validated Pipeline for 3D Segmentation of High Resolution Fluorescence Microscopy Images

3.1 Introduction

As highlighted in the general introduction, accurate and reliable anatomical segmentation of brain images is crucial for any study involving connectivity or function of the brain, as the conclusions of these studies are critically dependent on the correct assignment of anatomical regions to the neurons observed or manipulated in the experiment. With the advent of high-throughput whole-brain imaging, automated segmentation based on image registration has become essential to the field of rodent systems neuroscience (Kim et al., 2014b; Menegas et al., 2015; Renier et al., 2016; Vousden et al., 2015). At this moment however there remains a lack of quantification regarding the performance of automated segmentation on 3D fluorescence data.

3.1.1 Standardising Automated Segmentation

At this time we are lacking a standardised image registration pipeline for 3D fluorescence data of the rodent brain. Past studies either omit details about the software and parameters used for registration (Lein et al., 2007; Oh et al., 2014) or rely on unpublished in-house pipelines based on open source registration tools without full disclosure of the parameters used (Menegas et al., 2015; Vousden et al., 2015). Furthermore, while automated image segmentation is widely used in clinical research and the performance of many segmentation tools has been validated extensively on MRI data by comparison against the performance of human raters (Ou et al., 2014), such validation did not exist for 3D fluorescence data. There is hence a need for a fast, open and validated pipeline for automated segmentation of 3D fluorescence mouse brain data that could be used as a standardised segmentation tool by the community.

To accomplish this, I developed a package consisting of a precompiled version of NiftyReg (Modat et al., 2010), a registration toolkit for fast affine and b-spline-based free-form registration of MRI data (see 2.7.3), together with a custom user interface and full documentation. The atlas used to evaluate aMAP was by Kim et al. (2014b). It is a

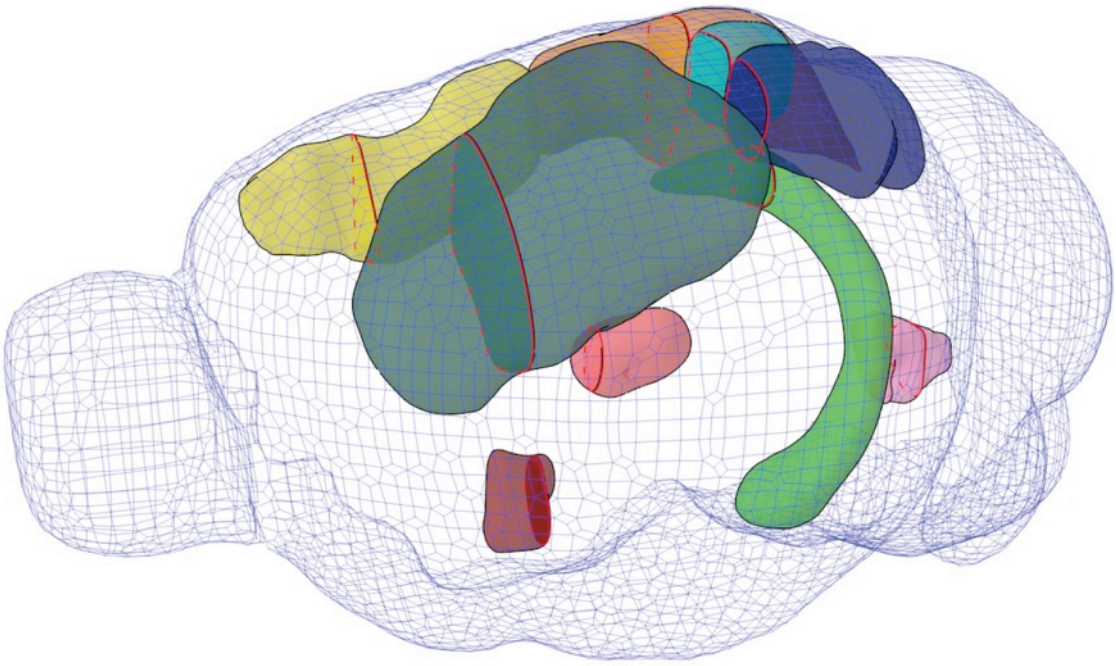


Figure 3-1: Illustration of the brain structures segmented by human raters and aMAP
An illustration of the 3D shape of the nine brain structures used to assess segmentation performance (left hemisphere). Red lines in each structure highlight the coronal plane of the reference atlas that was presented to human raters.

manually optimised atlas based on the original Allen atlas (Lein et al., 2007). For its use with aMAP, the structures in the Kim et al. atlas were smoothed along the z-axis to reduce the high-frequency noise in the segmentations along the z-axis (see 2.7.2). aMAP was capable of segmenting a complete STP dataset (downscaled to an isometric voxel size of 12.5µm, see 2.7.1) in 40 minutes on a Dell T7500 dual-processor workstation.

To validate aMAP-based segmentations, a cohort of 22 neuroscientists was split into two groups and each person was asked to segment 10 structures in 3 brains (Figure 3-1). The two groups were shown data from different brains, so a total of 6 brains were segmented, each by 11 raters. The segmentation quality of aMAP was then compared to that of the manual segmentations performed by human raters.

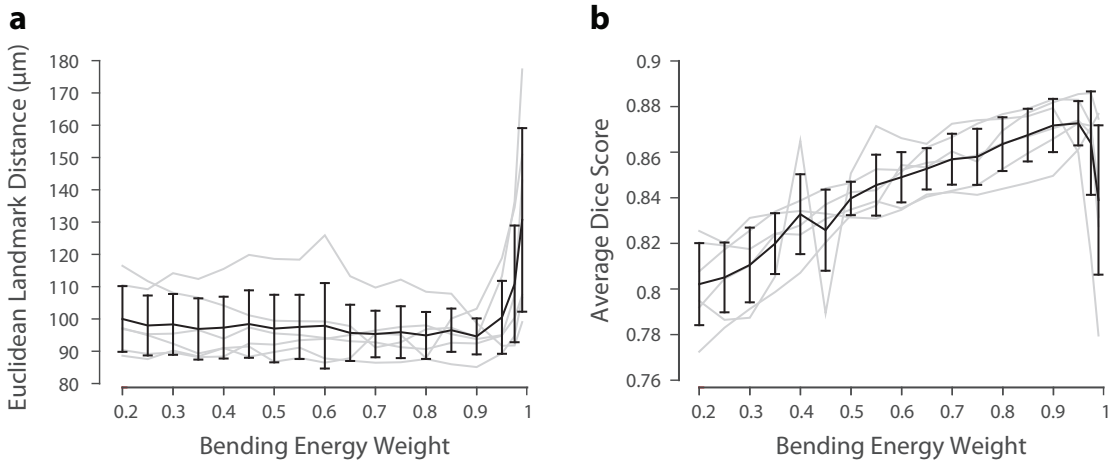


Figure 3-2: Sensitivity of ELD and STAPLE-Dice scoring

a) Plot showing mean ELD between ten standard landmarks in the reference atlas and its corresponding partners in the registered brain for all 6 brains used in this study, plotted against BE (grey lines, mean shown in black). b) Plot of Dice scores for the same brains and BE range (grey lines, mean shown in black). Error bars show the s.d.. BE: Bending Energy; ELD: Euclidean landmark Distance

3.2 Results

3.2.1 The Euclidean Landmark Distance Metric Does Not Accurately Report Registration Accuracy

Previous publications have used the Euclidean distance between anatomical landmarks (Euclidean landmark distance, ELD) to score registration accuracy, with the assumption that high registration accuracy equates to high segmentation accuracy (Kim et al., 2014b; Ragan et al., 2012). To calculate the ELD, anatomical landmarks are annotated in both, the reference brain and the individual STP datasets before registration. The Euclidean distances between the landmarks in the reference brain and the individual datasets are then measured after registration.

I evaluated the performance of ELD by generating free-form registrations of the Kim et al. reference brain to 6 individual STP brains with 18 different bending energy weights (BE) ranging from 0.2 to 1. BE controls the relative influence of the regularisation term on the registration score. Since the regularisation term penalises high-frequency

transformations of the image during registration, a low BE value will lead to overfitting while a too high BE will overly constrain the registration, resulting in global misalignment. Substantial changes in BE hence impact the quality of the resulting registration and any measure of registration quality should detect these changes. However, the ELD failed to report any significant changes in registration quality over a wide range of BE values (Figure 3-2 **a**, 0.2-0.95, repeated ANOVA, $F_{(17,75)}$=0.45, P=0.95).

3.2.2 Validation Using Manual Segmentations

The main issue when attempting to evaluate the quality of a segmentation is that there is no 'ground truth' to score against. To overcome this, several methods have been used in clinical research to generate a 'consensus segmentation' from multiple individual segmentations to be used as a ground truth estimate (Jorge Cardoso et al., 2013; Rohlfing and Maurer, 2007; Warfield et al., 2004). A commonly used method is simultaneous truth and performance level estimation (STAPLE, Warfield et al. (2004)), an iterative algorithm that simultaneously estimates the quality of multiple segmentations and the quality-weighted consensus segmentation.

To validate aMAP, the consensus segmentation was generated from the manual segmentations and used to assess segmentation quality of both, manual and automated segmentations. Segmentation quality was determined using the Dice score metric (Dice, 1945), which quantifies the overlap of two sets (in this case the consensus segmentation and the individual segmentation) and is widely used to score segmentation agreement (Jorge Cardoso et al., 2013; Klein et al., 2010; Leung et al., 2010; Modat et al., 2010; Nestor et al., 2013). As complimentary measures, consensus segmentations were also generated using shape-based averaging (Rohlfing and Maurer, 2007), and segmentations were further scored using the Hausdorff distance. However, as the results from these metrics were similar to those obtained from the more common STAPLE-Dice scoring, they are not discussed in detail (see Appendix, 6.1 and Methods).

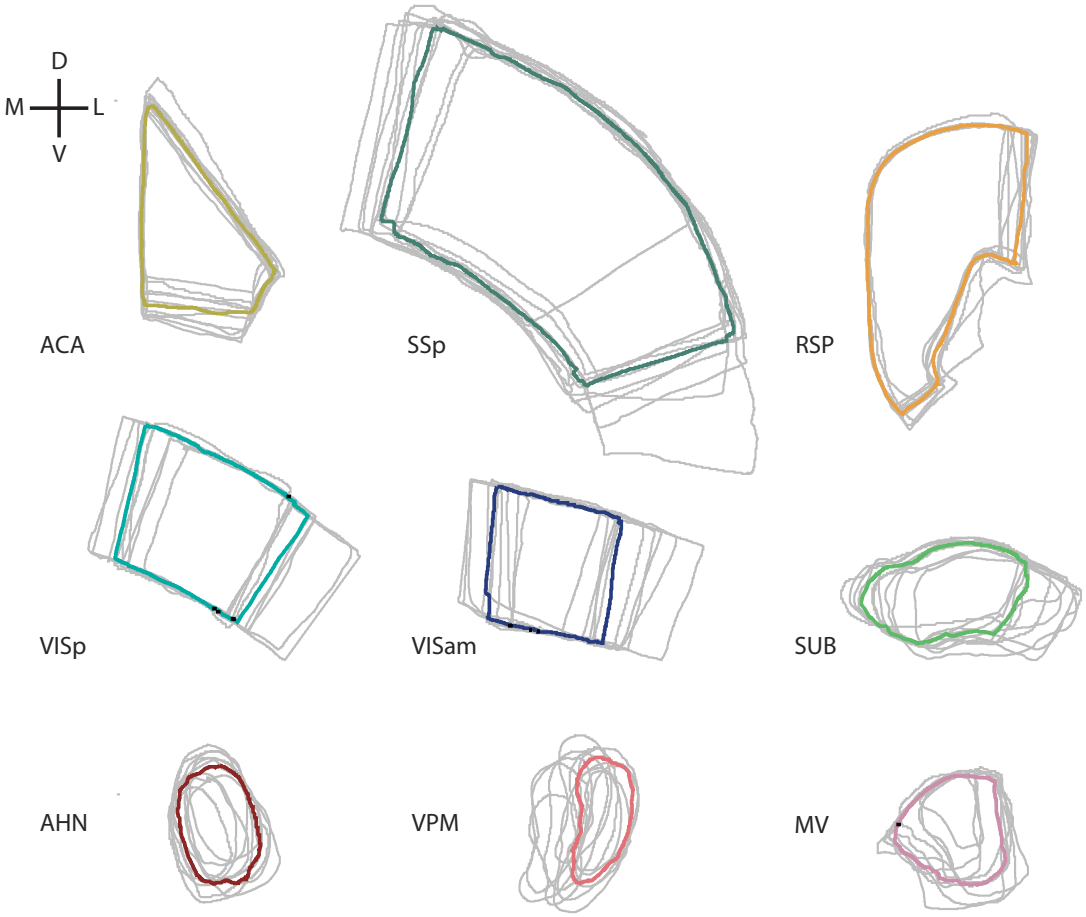


Figure 3-3: Outlines of segmentations performed by human raters

The segmentation outlines of all structures are shown for a group of 11 raters (grey lines). The STAPLE consensus segmentation for the same structures and raters is overlaid as a bold coloured line. ACA: anterior cingulate area; SSp: primary somatosensory cortex; RSP: retrosplenial cortex; VISp: primary visual cortex; VISam: secondary visual cortex, anteriomedial part; SUB: subiculum; AHN: anterior hypothalamic nucleus; VPM: ventral posteromedial nucleus of the thalamus; MV: Medial Vestibular Nucleus

To confirm the sensitivity of the STAPLE-Dice method to changes in segmentation quality, it was applied to the same datasets and BE range used to previously test the ELD method. While the ELD did not report changes in registration quality over a wide range of BE (Figure 3-2a, 0.2-0.95), the STAPLE-Dice method showed a clear increase in segmentation quality with rising BE values over the same range (Figure 3-2 **b**, repeated measures ANOVA, $F_{(15,75)}=16.8$, $P<0.001$).

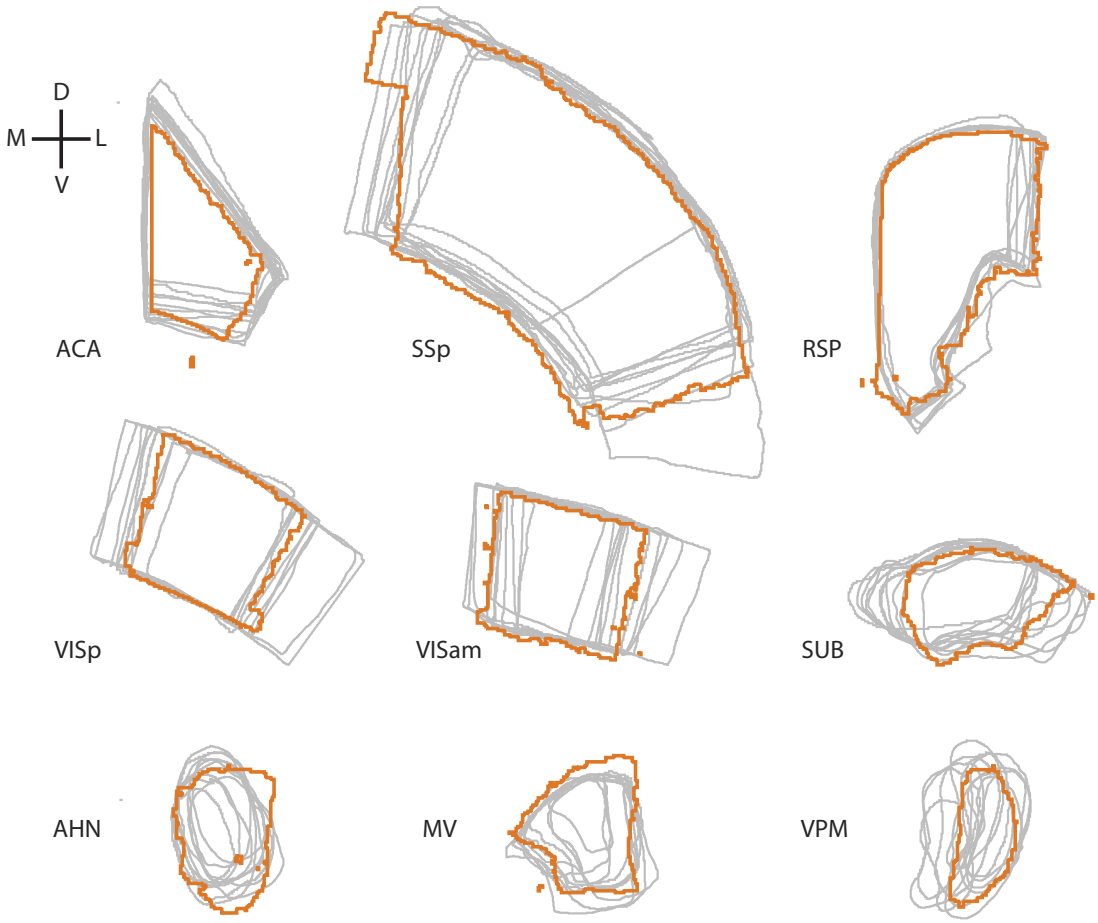


Figure 3-4: Outlines of aMAP and manual segmentation

Segmentation outlines of all structures are shown for a group of 11 raters (grey lines) with the corresponding aMAP segmentation result shown in orange. ACA: anterior cingulate area; SSp: primary somatosensory cortex; RSP: retrosplenial cortex; VISp: primary visual cortex; VISam: secondary visual cortex, anteriomedial part; SUB: subiculum; AHN: anterior hypothalamic nucleus; VPM: ventral posteromedial nucleus of the thalamus; MV: Medial vestibular nucleus

3.2.3 Qualitative Analysis of Manual and Automated Segmentations

While qualitative assessment of the segmentation outlines is arguably the most basic analysis of segmentation quality, it can nevertheless provide an informative visualisation of the variance between individual segmentations. The overlays clearly show that manual segmentations displayed substantial variability in the estimation of structure size and location between individual raters (Figure 3-3). This variability was especially apparent in structures that lacked easily identifiable anatomical borders, such

as the ventral posteromedial nucleus of the thalamus (VPM) or the medial/lateral boundaries of cortical structures. In general, there appeared to be higher agreement between raters on structures like the retrosplenial cortex (RSP) that could be identified using obvious anatomical landmarks.

Qualitatively, automated segmentations appeared to be similar to manual segmentations in location, size and shape. However, artefacts are noticeable in some segmentations (Figure 3-4, steps on the border of SSp and VISp, isolated blobs on ACA, RSP, VISam, AHN, MV).

3.2.4 Quantitative Comparison of Manual and Automated Segmentations

3.2.4.1 Median Performance

Analysis of the Dice scores confirmed the similarity between manual and aMAP segmentation. When pooling the scores from all brains and structures, there were no significant differences between the median scores of manual and aMAP segmentations (Figure 3-5 **a**, Mann-Whitney-U-test, median Dice of 0.91 vs. 0.92, P=0.52; n=4 brains, 9 structures, 22 human raters). Similarly, when grouping by anatomical structure there were no significant differences between manual and aMAP segmentations in 8 out of 9 structures, with humans scoring significantly better when segmenting the anterior cingulate area (Figure 3-5 **b**, ACA, Mann-Whitney-U-test, median Dice of 0.95 vs. 0.87, P=0.005). When grouping the Dice scores by brain instead of structure, there were no significant differences between the scores of automated and manual segmentations (Figure 3-5c, Mann-Whitney-U-test, P>0.49). However, while the median scores of manual and aMAP segmentations did not significantly differ, the scores of human raters displayed a substantial variance, while the performance of aMAP was significantly more consistent (Figure 3-5a, Levene's test on pooled scores of 4 brains, 9 structures and 22 human raters; s.d.: 0.16 vs. 0.05, p=0.005)

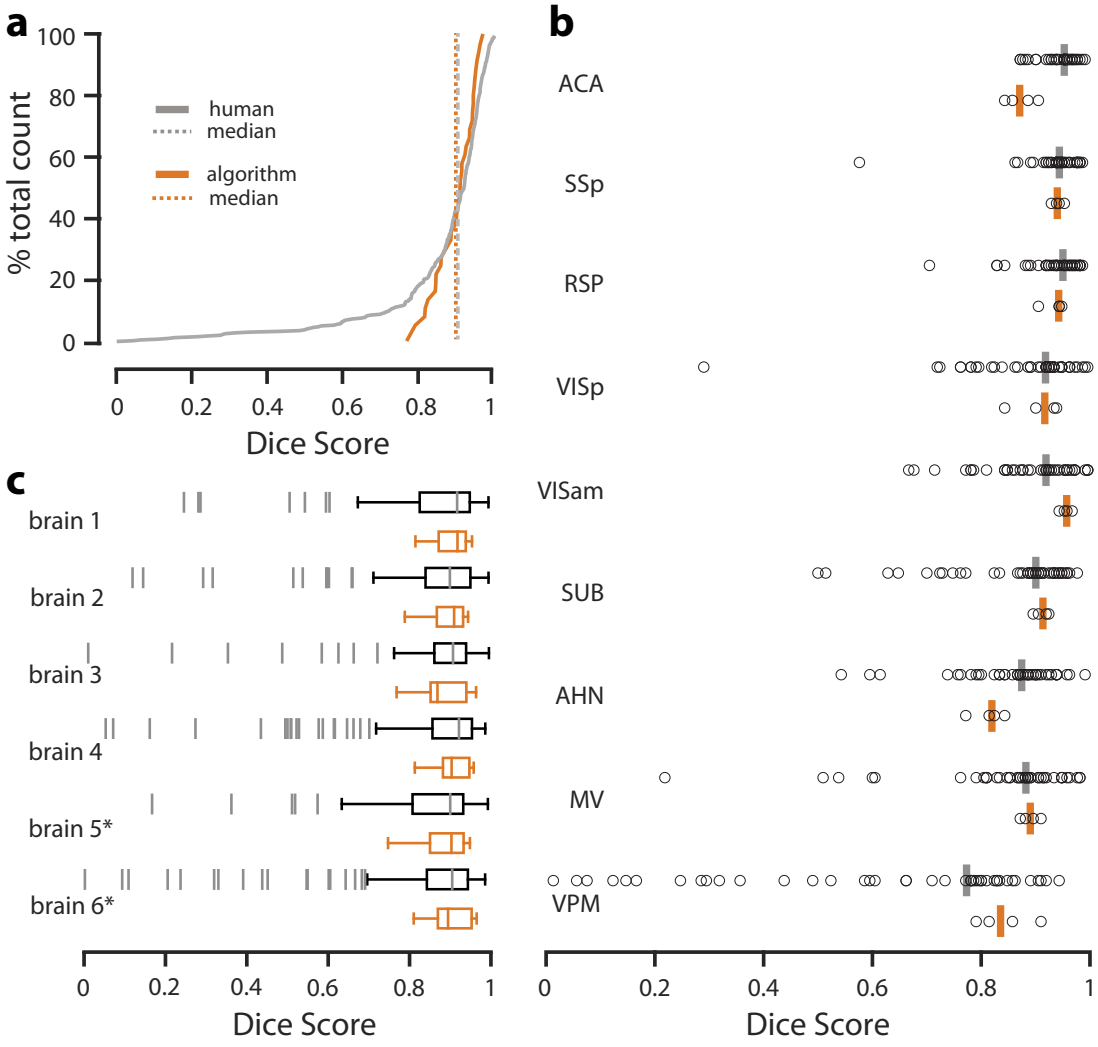


Figure 3-5: Dice scores of manual vs. aMAP segmentation

a) Cumulative histogram of the Dice scores for segmentations performed by human raters (grey, n=22 raters, each segmenting 9 structures in 2 out of 4 potential brains) and aMAP (orange, n=4 brains, 9 structures). b) The data from **a** grouped by structure. c) The data from **a** grouped by brain

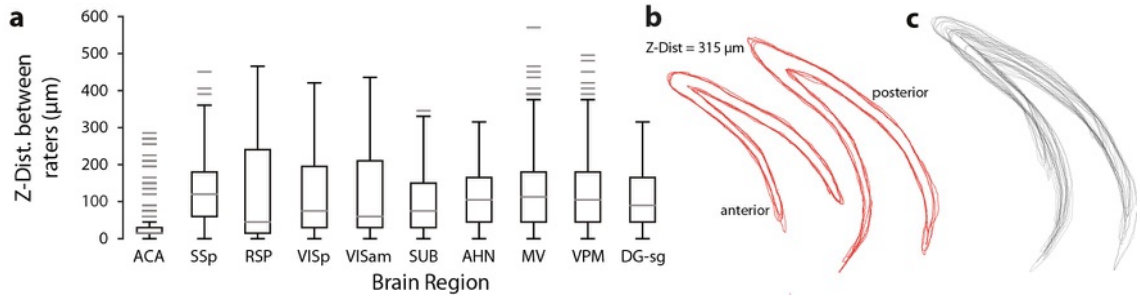


Figure 3-6: Disagreement in z-choice of human raters

a) Box plot showing the anterior-posterior distance between the estimation of the correct optical section (z-choice) of any two human raters (n=22 raters, each segmenting three of six potential brains). b) Outlines of two sets of 4 DG-sg segmentations from one brain, each set only including segmentations performed on the same plane. c) Outlines of all DG-sg segmentations performed on the brain used in **b**.

3.2.4.2 Sources of Variance in Human Rater Performance

In addition to the variability in the X-Y outlines, human raters also disagreed in their choice of the section that best corresponded to the presented Allen brain atlas images, resulting in substantial variations in the z-section chosen for segmentation (z-choice, Figure 3-6a). These differences in z-choice had an especially large influence on manual segmentations of the DG-sg since this structure changed substantially in shape over the presented z-range. As the DG-sg can be readily delineated in the STP datasets (see appendix 6.2), manual segmentations performed on the same optical plane showed a high degree of overlap (Figure 3-6b) while segmentations taken from different z-sections showed large x-y disagreement (Figure 3-6c). Since this x-y disagreement is unlikely to have been the result of uncertainty about the boundary of the structure, DG-sg segmentations were excluded from the analysis. To test whether any other structures were negatively influenced by the z-choice, all aMAP scores were compared with the scores of segmentations performed within 3 and 7 sections of each other. While the Dice score in SUB showed a significant increase when limiting the z-window, it did not result in a significant improvement compared to the aMAP score. Limiting the z-window did not lead to significant changes for any of the remaining brain structures (Figure 3-7).

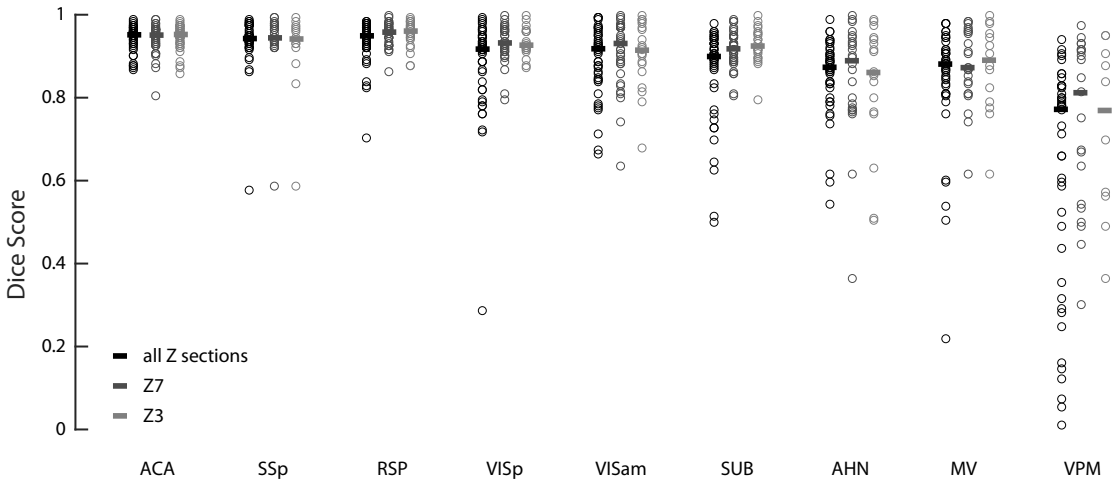


Figure 3-7: Z-window

Plot showing Dice scores for manual segmentations performed on all image planes (black), within 7 (Z7, dark grey) or 3 (Z3, light grey) optical sections of one another. New STAPLE consensus segmentations were calculated when limiting the z-window.

3.2.5 Reliability of Segmentation

While aMAP will always produce identical segmentations when applied to the same input data, this is not necessarily true for human raters. Hence, the inconsistency of human raters may be a substantial source of the observed variance in human Dice scores. To investigate the potential influence of rater reliability, all structures of one of the three brains were presented twice to each rater. The raters were unaware of this and repeat presentations of the same data were spaced at least 20 segmentation tasks apart (each rater completed a total of 40 segmentation tasks). The inter-rater Dice score was then obtained by calculating the overlap between a rater's first and second segmentation of the same structure (Figure 3-8 a).

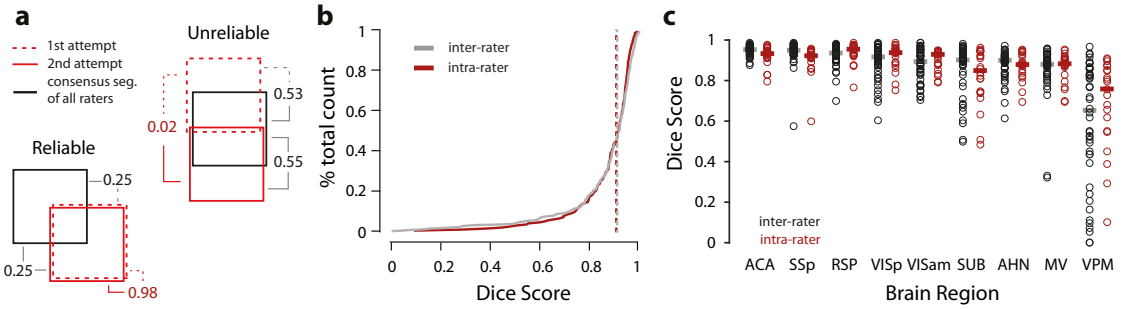


Figure 3-8: Rater reliability

a) Schematic highlighting two extreme segmentation reliability scenarios. Bottom left: When calculating the overlap with the consensus segmentation of all raters (black square), a rater may perform poorly (Dice score**:** 0.25, grey lines), yet be very reliable in their segmentation, as shown by the inter-rater Dice score (0.98, red). Top right: In contrast, a given rater may obtain a higher Dice score with the consensus segmentation (0.53 and 0.55, grey lines), but be unreliable in their estimate of the location of the structure (intra-rater Dice score**:** 0.02; red). b) Cumulative histogram of the intra-rater (red, n=22 raters, each segmenting 9 structures in 1 brain twice) and inter-rater (grey, n=22 raters, each segmenting 9 structures in 2 out of 4 potential brains) Dice scores for segmentations performed by human raters. c) Data from **b** grouped by structure.

Comparing the inter-rater Dice scores with the regular (intra-rater) Dice scores revealed no significant difference in overall median performance (Figure 3-8b Mann-Whitney-U-test inter-Dice, vs. intra-Dice: 0.92 vs. 0.91; P=0.32). When grouping by brain structures, inter-rater scores were significantly lower on the anterior cingulate area (ACA) and primary somatosensory area (SSp) (Figure 3-8c, Mann Whitney U-test, inter-Dice vs. intra-Dice: ACA: median 0.95 vs. 0.93, P=0.01; SSp: median 0.95 vs. 0.92, P=0.001) and not significantly different from intra-rater Dice scores on the remaining structures (Figure 3-8c P>0.21). The variance in inter-rater scores was reduced by a modest, but significant amount (Levene's Test, inter vs. intra: s.d.: 0.16 vs. 0.12, P=0.044). Taken together, the similarity of inter- and intra-rater Dice scores suggests that inconsistency in the segmentations of human raters is a major factor in the observed variance and disagreement between different raters.

3.3 Discussion

With the advancement of high-throughput 3D imaging methods and large-scale studies facilitated by these methods, fast accurate and reliable anatomical segmentation of data has become a critical issue. Since manual segmentation of whole brains is not feasible, automated segmentation has become a key element in the analysis of high-resolution 3D fluorescence data. Despite its widespread use, however, automated segmentation has not been validated against the segmentation performance of human raters on 3D fluorescence data of the mouse brain.

Previous publications have either reported qualitatively accurate segmentation without further quantification (Menegas et al., 2015; Renier et al., 2016; Vousden et al., 2015) or used ELD to quantify registration accuracy (Kim et al., 2014b; Ragan et al., 2012). However, the issues with ELD are twofold. Firstly, it is strictly a measure of registration accuracy and hence cannot directly report segmentation quality. Secondly, ELD can only report local registration accuracy at a set of individual points in the dataset. Therefore, its validity is dependent on whether the registration accuracy at these points is representative of the global registration accuracy. However, this is likely not true for anatomical landmarks, as these landmarks are located at (or close to) easily identifiable high-contrast areas. Since free-form image registration is non-linear and most image similarity measures are strongly influenced by high contrast areas, anatomical landmarks are likely to be correctly registered, even with parameters that cause errors in lower contrast areas. This explains why ELD was insensitive to overfitting and reported no changes in registration quality until the BE was high enough to cause a global registration mismatch. This is in line with the results of a previous study showing that ELD with sparse landmarks is unsuitable for free-form registrations of MRI data (Rohlfing, 2012). It is worth noting, that this finding does not invalidate the results of previous studies that used ELD to validate their automated segmentation pipeline, it merely shows that the accuracy of the underlying image registration cannot be judged based on the published result.

The preferred method to evaluate the suitability of automated segmentation tools in the clinical field is to directly compare their performance to that of human raters (Ou et al., 2014) and this study is, to my knowledge, the first to do so on fluorescence data of the mouse brain. On average, the quality of aMAP segmentations was on par with manual

segmentations of human raters. Interestingly, human raters showed substantial variance when performing the same segmentation twice, with the overlap between the first and second segmentation being no better (on average) than the overlap between the first segmentation and the agreement segmentation of all raters. In contrast, aMAP is purely deterministic and performed reliably, with a significantly lower overall variance in segmentation scores compared to human raters.

It is worth noting that ideal conditions were provided for manual segmentation by ensuring correct coronal alignment of all STP data (by rigid registration to the STP reference brain of the Allen atlas). In addition, the z-plate of the atlas to be used for each segmentation was predefined. It is thus possible that disagreement between human raters may be even larger in a "real-world" experimental scenario where data might not be optimally aligned and differences in perceived z-position may cause raters to base their segmentation on different atlas plates.

Despite its high reliability, automated segmentation using aMAP is inherently dependent on similarities between the reference brain and the individual datasets. To maximise the similarity, it is recommended to perform registration on the background fluorescence channel, as strong staining patterns may negatively influence the registration process in a way that is difficult to predict. Similarly, strong artefacts in the STP data, such as dissection damage, uneven background illumination or bright patterns, as caused by a failed perfusion, can negatively affect the quality of automated segmentation. It is thus advisable to perform a manual quality control on all segmentations, for example by overlaying the segmentation outlines with the original STP data.

Ultimately, automated segmentation has the potential to greatly extend the usefulness of the studies that employ it, since it allows to publish locations of interest, be it cells or activity patterns, in the coordinate-space of a reference atlas. By doing so, any future updates or improvements of that atlas can immediately be applied to the published data, ensuring that the data remains accurate and relevant. To ensure comparability between studies, it would be highly desirable to establish a standardised procedure for automated segmentation. Due to its accuracy, reliability and high speed, I believe that aMAP is ideally suited to become the standard tool for mouse brain segmentation.

Chapter 4. Automated Neuron Detection Using the Fast Automated Cell Counting Tool (FACCT)

4.1 Introduction

While attempts to automate the identification and localisation of individual cells in microscopy data are not limited to the field of brain mapping or even neuroscience, the recent developments in whole-brain 3D fluorescence imaging methods represent both a strong incentive and a unique challenge to automated cell detection. Large whole-brain datasets can now be acquired at high speed (~1TB/day, ~2.5TB/brain), which severely limits the number of feasible cell detection algorithms and requires a stronger focus on computational optimisation. At the same time, the high variability of background structures in a dataset encompassing an entire brain presents a unique challenge to automated cell detection.

While a number of 3D cell detection algorithms have been introduced recently, they were mostly applied to small high-resolution datasets, on the assumption that all structures with a certain morphology and size will be labelled and identified as cells (He et al., 2015b; Oberlaender et al., 2009; Toyoshima et al., 2016). This makes them unsuited for our STP datasets, which are of comparably low resolution, yet contain artefacts that resemble cells morphologically. In addition, the large size of our datasets would require several weeks of processing time per brain, even when assuming linear scalability of these algorithms, which may not be the case with very large datasets.

So far, whole-brain cell detection has been mostly performed with in-house pipelines, making a direct comparison of their performance difficult (Kim et al., 2014b; Menegas et al., 2015; Vousden et al., 2015). This issue is further compounded by the fact that quantification of the performance is either completely lacking (Menegas et al., 2015; Vousden et al., 2015) or has only been carried out on small subsets of a whole-brain dataset (Kim et al., 2014b; Renier et al., 2016).

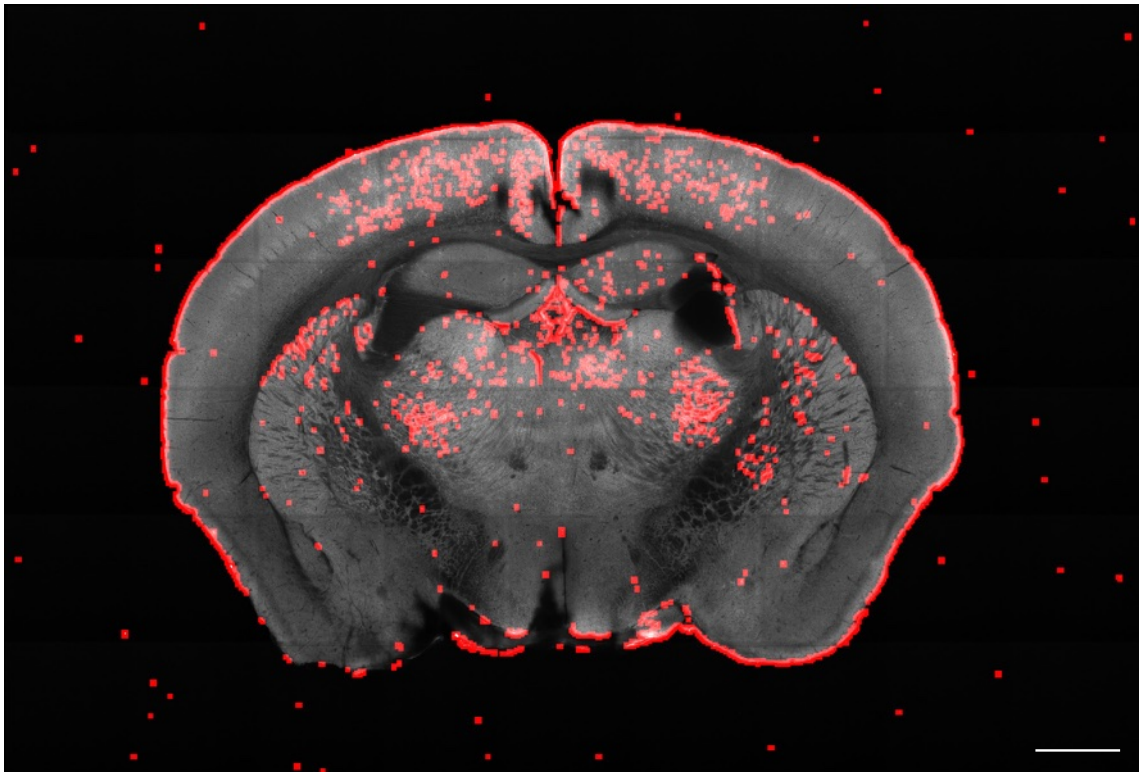


Figure 4-1: Tile classifier

Illustration of the areas marked by the tile classifier for further analysis (highlighted in red). Coronal view of the brain of an adult Ntsr1-Cre mouse injected with rabies virus into the primary visual cortex and imaged using STP. Scale bar: 1mm

This chapter therefore focuses on the development of a fast automated cell counting tool (FACCT), designed to rapidly and accurately detect fluorescently labelled cells in whole-brain STP datasets. It uses a combination of custom filters to find centroid structures within a certain size window as potential cell locations, which are then classified by a deep learning module to remove false positives. To ensure high processing speed, the initial filters are realised using a highly optimised custom image processing framework that integrates with ImageJ (Schneider et al., 2012) and allows easy development of fast, multithreaded image analysis routines.

4.2 Results

4.2.1 Tile Classifier

To cope with uneven illumination, FACCT splits the dataset into small overlapping square tiles, with an edge length of twice the soma diameter of the average expected cell (2x15µm). The tile classifier is the first step in cell detection and marks the areas of the dataset that contain relevant levels of signal by calculating the difference in brightness between the noise floor and the brightest elements for each tile. This difference is then compared to a threshold that is either set by the user or calculated automatically (see 2.11.6.1). Any tile where the brightness difference is smaller than the threshold is discarded from further analysis.

The advantages of this pre-classification are twofold: firstly, the process is relatively fast (under 5 hours per STP dataset) and reduces the amount of data that needs to be processed by the remaining filters by at least an order of magnitude. Secondly, it allows to design the remaining filters under the assumption that all input data could potentially contain cells (Figure 4-1).

4.2.2 Thresholding

The goal of the thresholding step is to separate the signal from the background in the image. This is particularly challenging as brightness levels can change significantly, both between and within an STP dataset. A thresholding algorithm for such data must therefore calculate thresholds on a local level and be capable of processing both very dark and bright areas, ideally without relying on external correction factors that may need to be "tuned" for each image. Unfortunately, in a qualitative visual evaluation, no tested pre-existing local thresholding paradigm was capable of performing reliably in both, high and low signal-to-noise scenarios (Figure 4-2). Hence, a custom thresholding algorithm was developed that combines the histogram-based analysis of the Otsu thresholding algorithm (Otsu, 1975) with an analysis of the resulting structures (see 2.11.6.2). Briefly, Otsu's algorithm finds the threshold that minimizes the intra-class (brightness) variance for both, thresholded and non-thresholded areas. In STP data, this generally leads to good results on areas with cells, but can result in under-estimation of the threshold in areas containing neurites or large background artefacts. To solve this

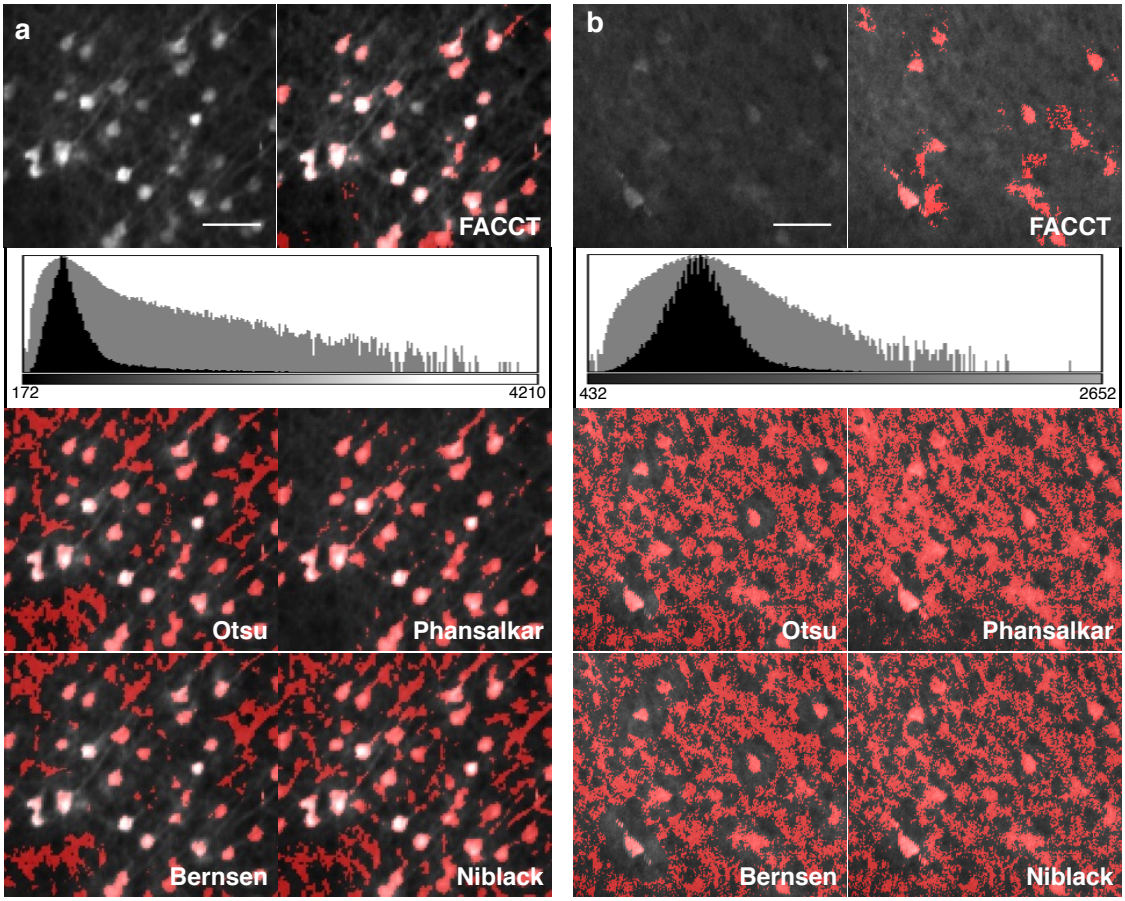


Figure 4-2: Comparison of thresholding methods

a) Panels showing the results of various local thresholding algorithms in a high-contrast example. Thresholded areas are marked in red. The grey area in the histogram shows the data on a logarithmic scale, the values indicate the min/max brightness levels. b) Panels showing the performance of the same thresholding algorithms in a low-contrast example. All images show views of the cortex of adult Ntsr1-Cre mice injected with rabies virus into the primary visual cortex and imaged using STP. Scale bars: 50µm

issue, the structures resulting from the Otsu thresholding are analysed. If the largest structure is too big to be a cell, Otsu's threshold is recalculated from the sub-histogram above the previously determined threshold. This process is repeated until the thresholding results in a structure that falls within the size expected for a cell. This "size-checked" local Otsu variant, in combination with the tile classifier, performed well in qualitative checks and was hence used for the thresholding module (Figure 4-2).

4.2.3 3D Soma Filter

While the cells in the images are marked as a result of thresholding, these thresholded areas also contain a large number of non-somatic structures, such as dendrites and noise (Figure 4-3b, blue areas). In a next step, the data is hence filtered using a "virtual nucleus" to determine the location of the somata and remove any remaining small structures. To achieve this, a spheroid is moved through the thresholded parts of the data. All locations where the spheroid overlaps with the thresholded areas by more than a certain percentage (default: 90%) are marked (Figure 4-3**b**, red areas). The marked areas are then size-filtered (default: 10 voxels), counted and their centres of mass are stored as potential cell positions (Figure 4-3**c**, red dots). The 3D soma filter not only removes small structures, but also separates clustered cells with touching perimeters (e.g. neurites) that would otherwise be considered as a single structure.
Finally, the brain is segmented using aMAP (see Chapter 3) and all marked locations outside of the brain are discarded.

4.2.4 Qualitative Analysis

To evaluate the performance of the filter pipeline, it was applied to STP data of the brain of an adult transgenic mouse expressing Cre under the control of the Ntsr1 promoter. Cells were labelled by injection of Cre-dependent rAAV expressing TVA and RG, followed by injection of RV expressing mCherry into the primary visual cortex. The fluorescent cells in the STP dataset were then marked by an expert human rater. Qualitative assessment of the locations marked by FACCT revealed that the software appeared to accurately detect the cells in the dataset, however the data also contained a large number of false positives, especially in bright and noisy areas around the olfactory bulb and the surface of the brain (Figure 4-4). In total, the human rater marked 19507 cells, while FACCT marked 58494 locations, of which 17520 had no human marked cell within a radius of 100µm (~30%).

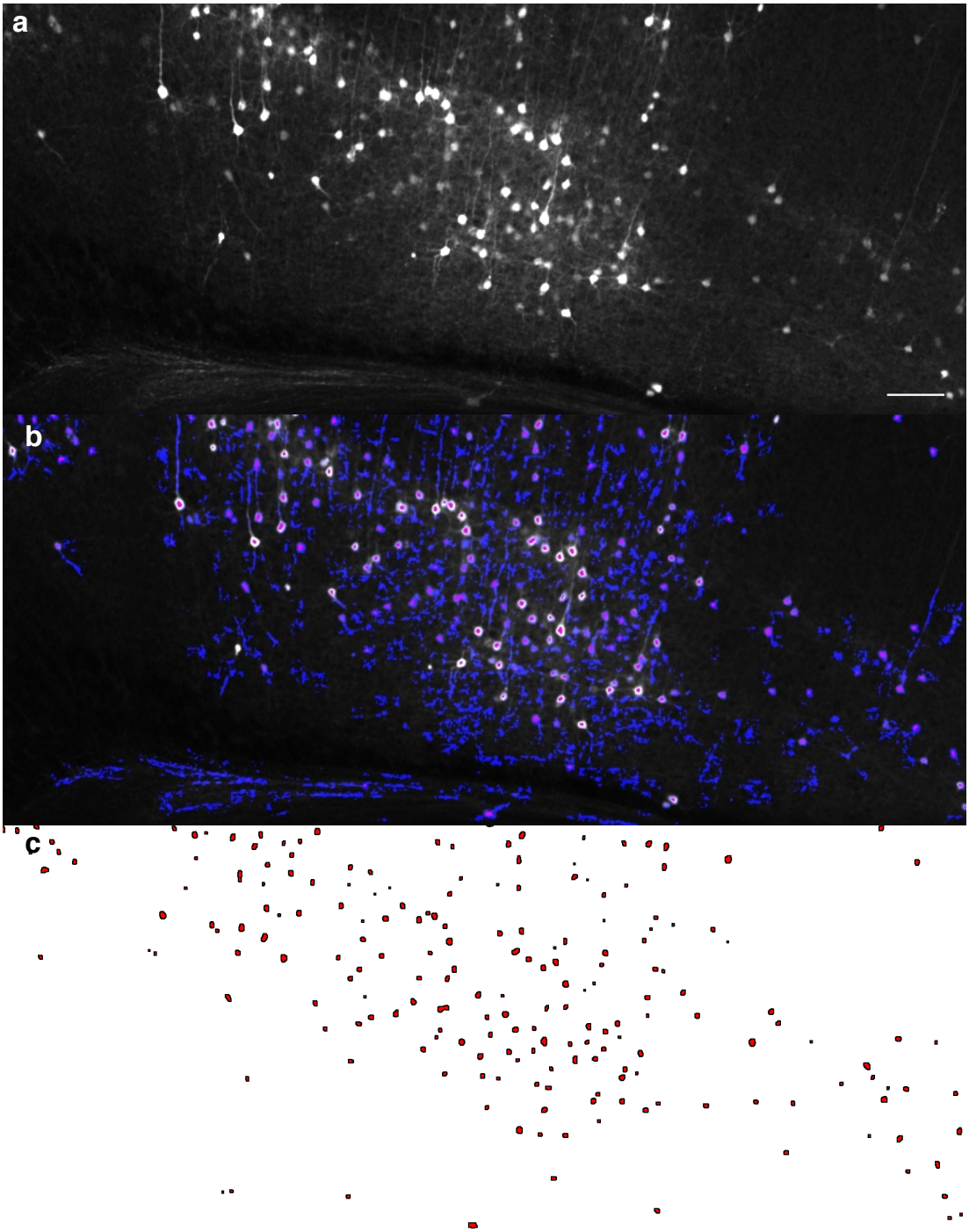


Figure 4-3: Result of thresholding and nucleus detection

a) Coronal view of the cortex of an adult Ntsr1-Cre mouse injected with rabies virus into the primary visual cortex and imaged using STP. b) Overlay of the result of thresholding and 3D soma filtering in the area shown in **a**. Thresholded areas are shown in blue, areas marked by the soma filter are highlighted in red. c) Areas marked by the soma filter highlighted in red. Scale bar: 100µm

4.2.5 Classification Using Deep Learning

These results show that a more detailed analysis of the STP data is required to accurately separate neuronal somata from spheroid non-cellular elements. Deep learning methods have recently matured and are now considered the state-of-the art in automated image analysis, outperforming all other approaches in image classification and object detection tasks to date (Russakovsky et al., 2015).

While deep learning is most widely used in 2D image classification tasks, some toolkits now allow processing of multidimensional data, making it possible to apply deep artificial neuronal networks directly to STP data. FACCT was hence complemented by an artificial neuronal network with a residual network architecture (ResNet) implemented in the deep learning framework Caffe (Jia et al., 2014) (see 2.11.7). The ResNet was chosen due to its high classification performance and good trainability (He et al., 2015a). Cuboids of 50x50x20 voxels centred on the locations labelled as potential cells by the previous filters were used as input. The network was given all three colour channels as input and trained to identify cells in the red channel, which contained the signal of the RV. The full data of three brains (Figure 4-7 & Figure 4-8, brains 4-6) and partial data of further three brains (Figure 4-7 & Figure 4-8, brains 1-3) were used for training (see 2.11.7).

Comparing the results before (Figure 4-4) and after the addition of the deep learning module (Figure 4-5) clearly showed the reduction of false positives. After filtering with the ResNet, 39829 (out of 58494) locations remained, of which now 827 had no human marked cell in a radius of 100µm (~2% vs. ~30%). However, the number of cells marked by FACCT was still roughly twice as high as that of a human rater.

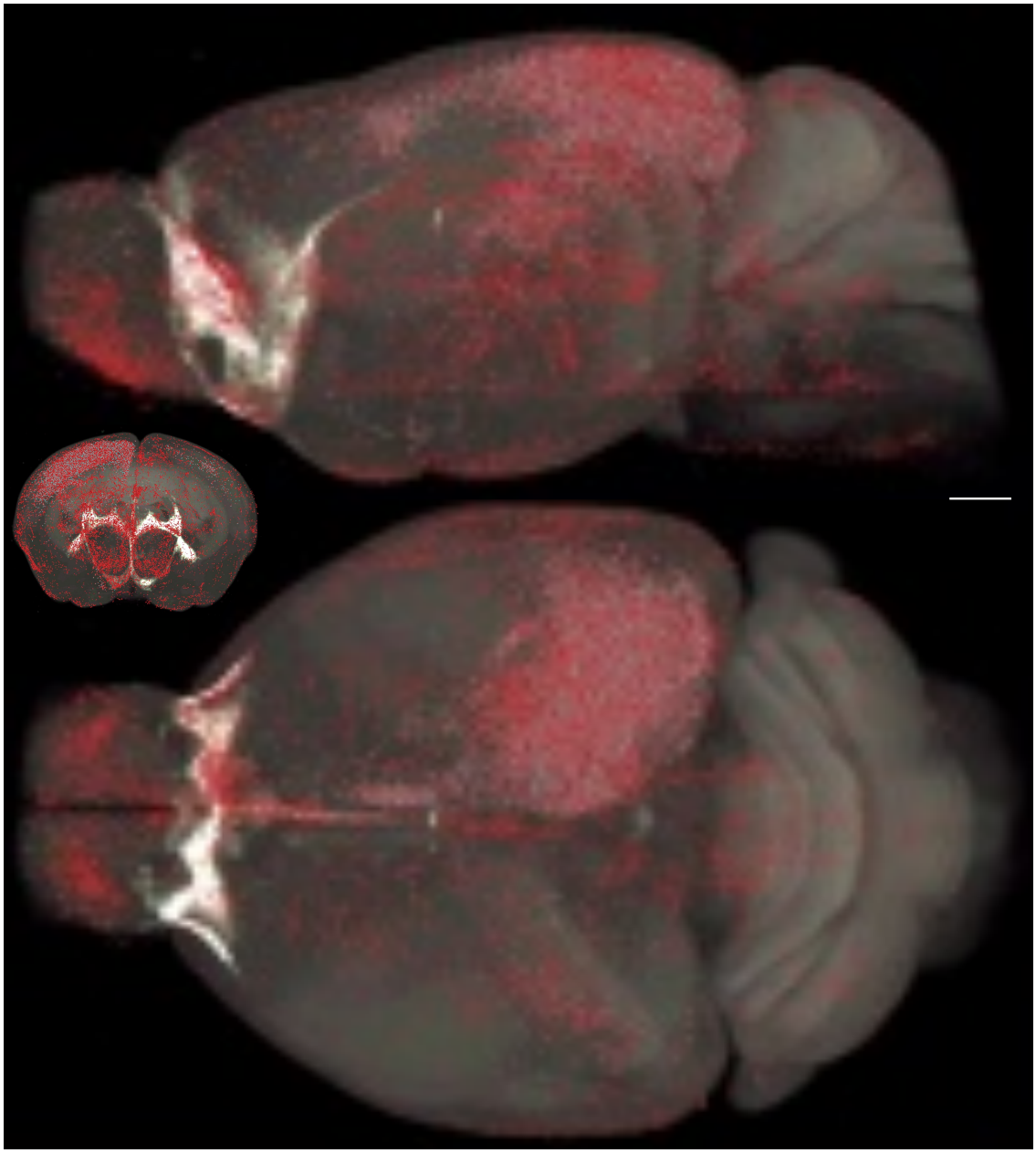


Figure 4-4: Thresholding and nucleus detection in a complete brain

Visualisation of the cells detected by a human rater (cyan) and FACCT (red) in a full STP dataset. The cells are shown on a maximum intensity projection of the Allen average brain, transformed to match the original STP dataset to provide a spatial reference. Top: sagittal view; Inset: coronal view; Bottom: horizontal view. Scale bar: 1mm

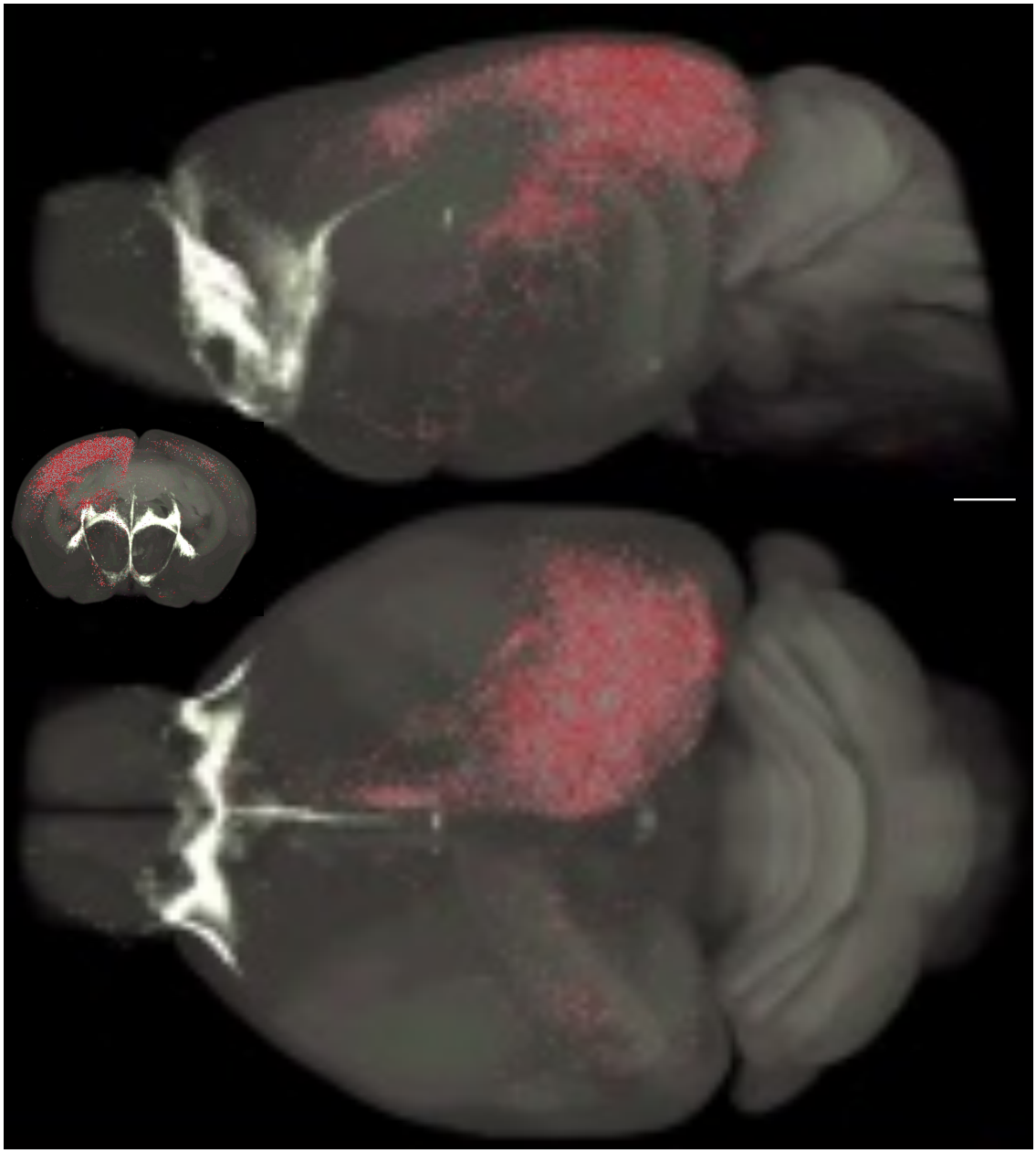


Figure 4-5: Whole-brain cell detection with added deep learning module

Visualisation of the cells detected by FACCT after complementation with a deep learning module (red) on the same STP dataset used in Figure 4-4. Cells marked by a human rater are shown in cyan. The cells are shown on a maximum intensity projection of the Allen average brain transformed to match the original STP dataset to provide a spatial reference. Top: sagittal view; Inset: coronal view; Bottom: horizontal view. Scale bar: 1mm

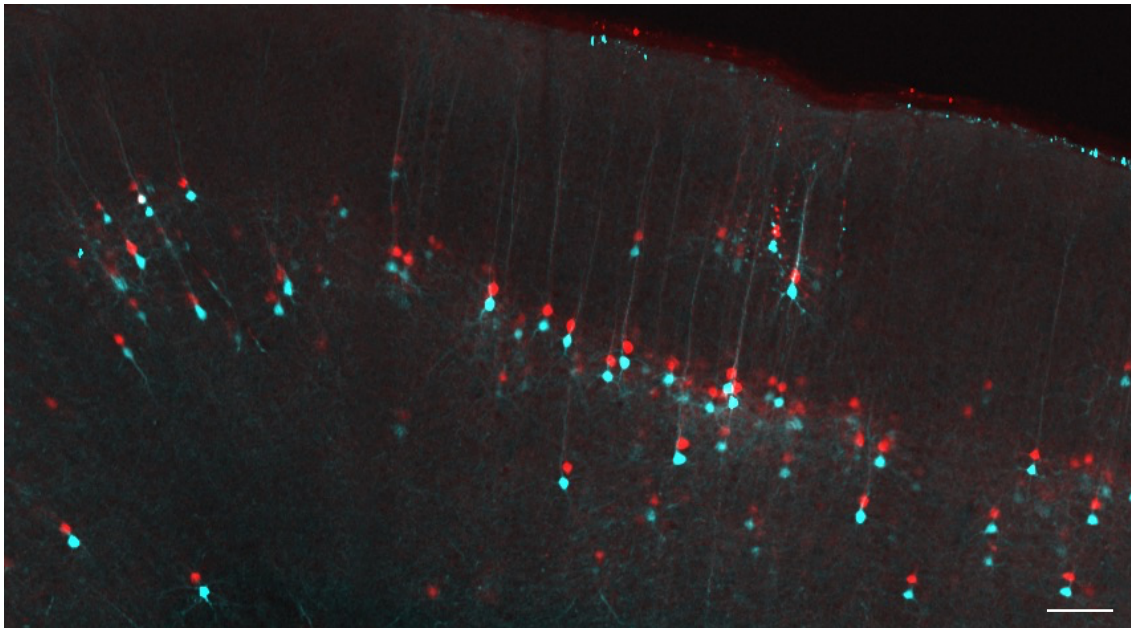


Figure 4-6: Example of z-discontinuity

Overlay of two consecutive STP images at the border from one physical section (cyan) to the next (red), showing a pronounced shift in z. Scale bar: 100µm

4.2.6 Multi-Labelling of Cells

The main reason for the overestimation of cell numbers by FACCT was an issue with the data from the STP microscope. To cover the complete surface of the brain, the microscope scans multiple overlapping tiles in a mosaic pattern. Unfortunately, the x/y coordinates reported for the individual tiles are inaccurate, resulting in a mismatch at the border region of individual tiles. The software that assembles the full images from the tiles attempts to correct the tile position in x and y, but does not ensure consistency between individual physical sections (along the z-axis). As a result, there are noticeable shifts in the images between individual physical sections (Figure 4-6). These shifts can occur every 50µm along the z-axis and as a result, the majority of cells were detected more than once. Because each of these multiple detections was located on a soma, the ResNet accurately recognised them as a cell and hence did not remove them from the dataset.

4.2.7 Quantitative Analysis

4.2.7.1 Count Comparison Using Equal Sampling

To quantify differences between the cells marked by FACCT and those of a human rater, the brains were segmented into equal cubes of 400µm edge length at 200µm intervals and the numbers of cells per region detected by FACCT were compared to the numbers of cells marked by the rater. As there are differences in the specified location of a detected cell between a human rater and FACCT, further compounded by the z-discontinuities in the data (as mentioned above), the cube size was chosen to be large enough to limit the influence of errors at the border of the cube.

The quantification of FACCTs performance was carried out on the data from the brain shown in Figure 4-4 and Figure 4-5 (Figure 4-7, brain 1) and 5 further brains, all injected with rAAV/RV into the primary visual cortex (Figure 4-7, brains 2-6). Without the ResNet, FACCT detected a large number of false positive cells (Figure 4-7 a; large sum of squared errors (SSE), high number of cubes with no human counts and high FACCT counts). Furthermore, while the overall correlation between human counts and FACCT counts was high (0.82-0.95), the slope was greater than one, showing that FACCT routinely overestimated the number of cells (Figure 4-7 a, slope 1.26-1.97).

When analysing the performance of FACCT after the addition of the ResNet, it became apparent that the ResNet reduced the large number of false positives, leading to a significant reduction of the sum of squared errors (64807->15842, 40281->9517, 54224->11180; one-tailed paired t-test, $p=0.008$) and a lower number of areas containing only FACCT counts. Despite that, the slopes of the regression line remained extremely similar (1.97 -> 1.91, 1.78 -> 1.74, 1.26-> 1.24), suggesting that the deep learning did not have a significant influence on the multiple-detection events. Furthermore, the slopes varied between individual experiments, meaning that the severity of the z-discontinuities differed between individual brains.

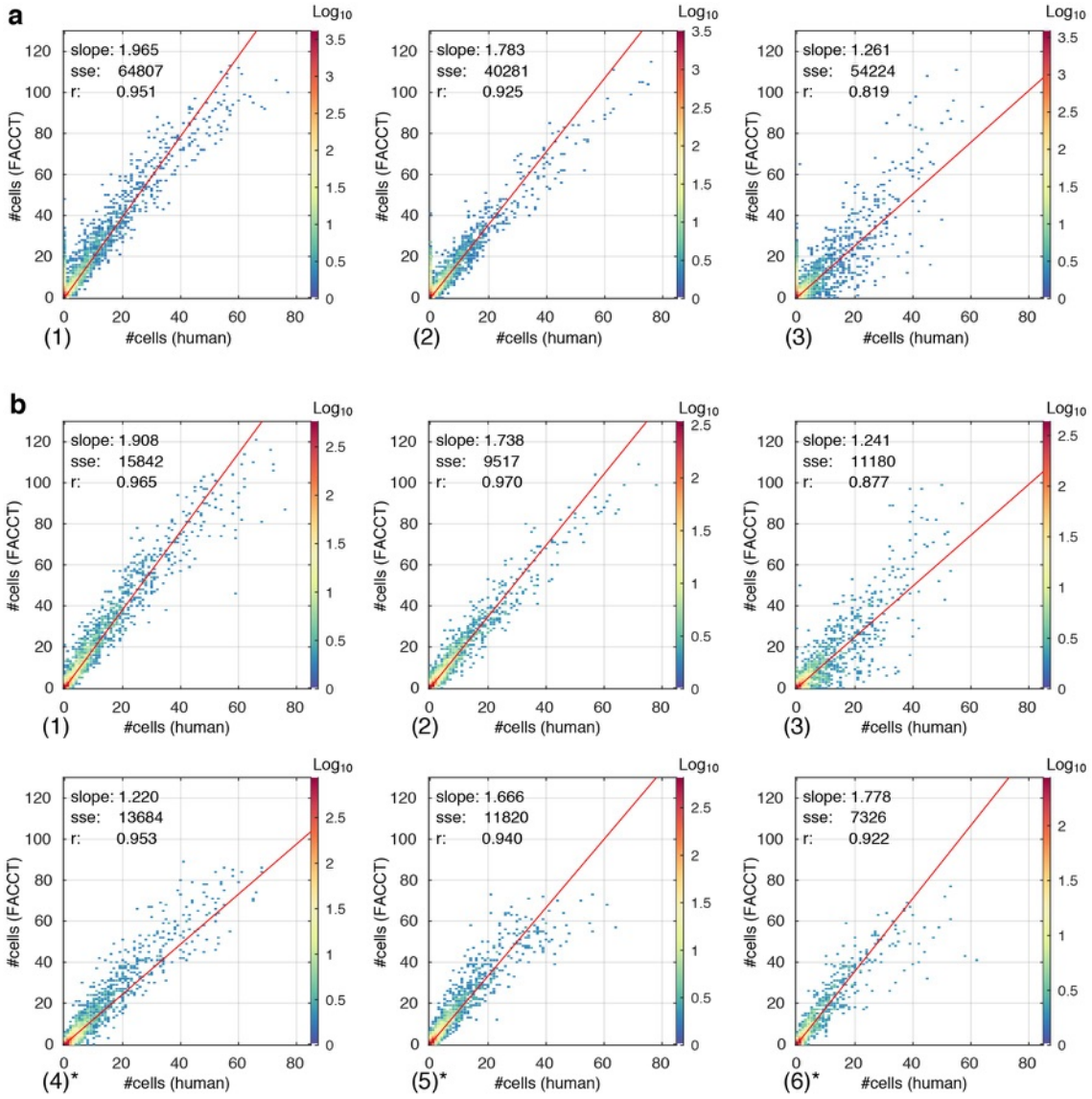


Figure 4-7: Comparison of FACCT and human cell counts

2D scatter plot showing the number of cells found by FACCT plotted against the number of cells found by a human rater. Cells were counted in STP data of mice injected into the primary visual cortex with RV. Each point represents a cube with an edge length of 400µm. Cubes were evenly spaced at 200µm intervals. Each histogram represents data from one brain. a) FACCT vs. human counts before deep learning b) FACCT vs. human counts after complementation with a deep learning module. Brains marked with an asterisk were used as training data for the ResNet. SSE: sum of squared errors; r: correlation

4.2.7.2 Count Comparison Using Anatomical Structures

4.2.7.2.1 Grouped by Brain

As the ultimate goal of the data analysis pipeline is to detect and anatomically map the individual cells, the STP data was segmented using aMAP with the Allen Common Coordinate Framework atlas (CCF, October 2016 release) to investigate whether segmentation by anatomical regions rather than regularly spaced cubes had any significant influence on the results. Plotting the number of cells detected by FACCT versus those marked by a human per brain region further confirmed the results from the cube segmentations (Figure 4-8). Slope and correlation were similar to the data from cube segmentations and the addition of the ResNet also significantly reduced the SSE (166724 -> 14928, 157464 -> 5631, 135246 -> 8860; p=0.002). Interestingly, while the results before filtering with the ResNet showed a lower correlation and higher SSE (Figure 4-7 a, Figure 4-8 a), the opposite is true after addition of the ResNet (Figure 4-7, Figure 4-8). This further supports the hypothesis that the increased number of cells reported by FACCT was due to multiple detection of existing cells rather than stochastic noise.

4.2.7.2.2 Grouped by Anatomical Region

To investigate whether FACCT's performance depended on the brain region, the data was next grouped by anatomical structure in a top-down approach. To account for the systematic overestimation of cells by FACCT, the cell counts were not analysed in terms of absolute cell numbers, but rather as a percentage of the total number of cells for each brain. The percentages reported by FACCT and human raters were then compared for each brain region.

When using a high-level segmentation (Figure 4-9), the regions that displayed high variability in their errors were (from lowest to highest) hippocampal formation, fiber tracts, brain stem and isocortex. The data from these regions were hence further analysed to identify the sources of variability.

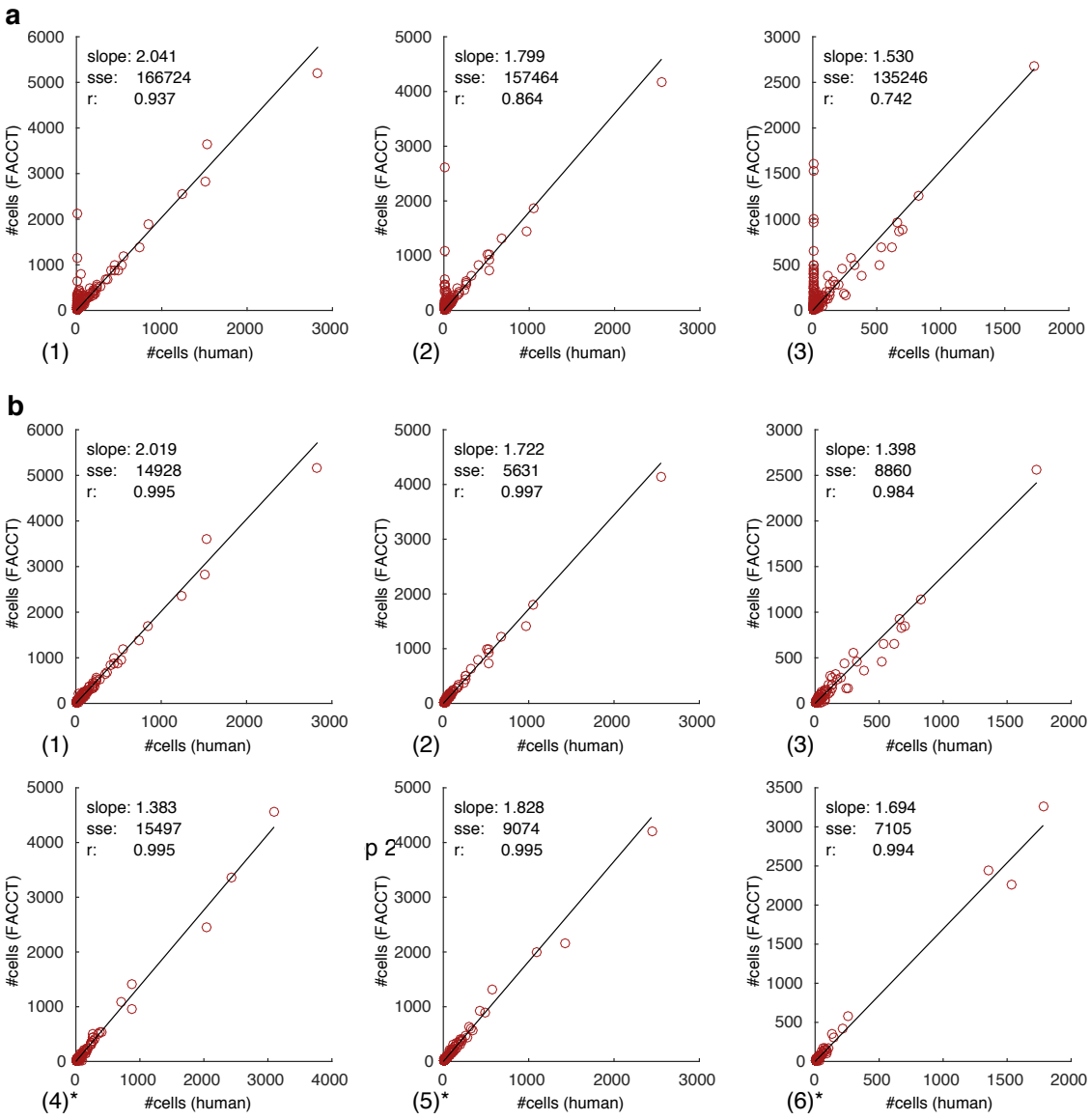


Figure 4-8: FACCT vs. human cell counts by region

Scatter plots showing the number of cells found by FACCT in a particular brain region plotted against the number of cells marked by a human rater. Each plot represents a brain and each dot represents a brain region. Brains are numbered in accordance with Figure 4-7. a) FACCT vs. human counts before deep learning b) FACCT vs. human counts after complementation with a deep learning module. Brains marked with an asterisk were used as training data for the ResNet. SSE: sum of squared errors; r: correlation

The hippocampal formation showed generally low disparity between human and automated counts, except for one outlier per region. Indeed, all positive outliers occurred in one STP dataset and all negative outliers in another. Both datasets exhibited numerous artefacts, such as shadows (most likely caused by pieces of floating dura) and bright specks of autofluorescence around the hippocampal regions that negatively impacted FACCT's performance in that area.

The high number of counting errors in the fiber tracts is surprising, as these areas should, by definition, not contain a large number of neurons. However, the neurons detected here are mostly the result of slight inaccuracies in the automated segmentation, causing the boundary of these tracts to be extended. In line with that, the area with significant errors was the corpus callosum, which was due to a slight dorsal shift of its boundary into the bottom of layer 6 of the cortical areas near the viral injection site (VISp). The cause underlying variability in cell detection was hence the same as in the cortical areas (see below).

The areas in the brainstem containing a noticeable variance in counting error were the dorsal part of the lateral geniculate complex (LGd), thalamus (TH) and lateral dorsal nucleus of the thalamus (LD). For the LGd and LD, this can at least be partly attributed to bright specks in some of the datasets that can be difficult to distinguish from cells by both, FACCT and human raters.

When evaluating the performance in cortical areas, the first notable result is a single dataset with severe underestimation of cell numbers in the retrosplenial area. This is due to a strong imaging artefact, most likely caused by a piece of uncut dura partially obstructing the objective at the retrosplenial cortex of that particular brain. As a result, FACCT was unable to accurately detect cells in that part of the STP dataset. The second prominent result is the high variance in the visual areas.

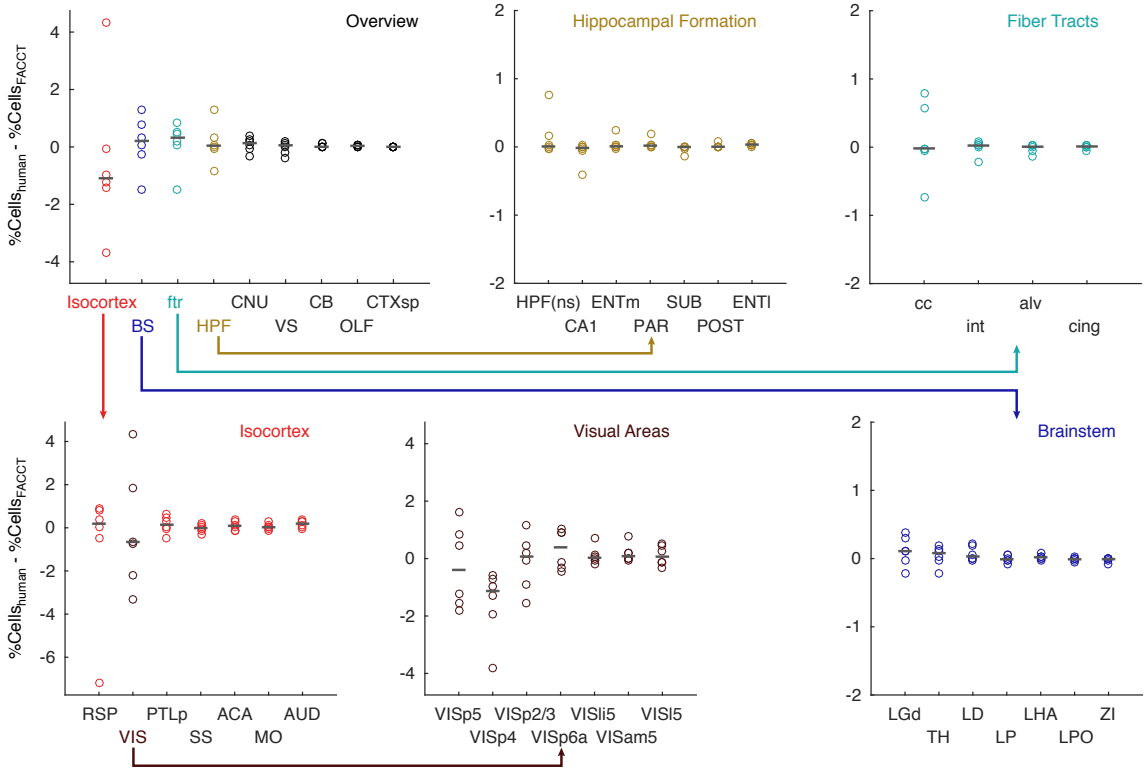


Figure 4-9: Performance of FACCT grouped per brain structure

Scatter plots showing the difference between human and FACCT cell detection grouped by brain region for the 6 brains shown in Figure 4-8. To correct for the systematic overestimation of cells by FACCT, data are compared as percentage of total cells. Plotted regions are sorted by the range of their values. BS: brain stem; ftr: fiber tracts; HPF: hippocampal formation; CNU: cerebral nuclei; VS: ventricular systems; CB: cerebellum; OLF: olfactory areas; CTXsp: cortical subplate; HPF(ns): hippocampal formation, no substructure; CA1: cornu ammonis, area 1; ENTm: entorhinal area, medial part, dorsal zone; PAR: parasubiculum; SUB: subiculum; POST: postsubiculum; ENTl: entorhinal area, lateral part; PRE: presubiculum; cc: corpus callosum; int: internal capsule; alv: alveus; cing: cingulum bundle; RSP: retrosplenial area; VIS: visual areas; PTLp: posterior parietal association areas; SS: somatosensory areas; ACA: anterior cingulate area; MO: somatomotor areas; AUD: auditory areas; VISp: primary visual area, layer 5; VISp4: primary visual area, layer 4; VISp2/3: primary visual area, layer 2/3; VISp6a: primary visual area, layer 6a; VISli5: laterointermediate area, layer 5; VISam5: anteromedial visual area, layer 5; VISl5: lateral visual area, layer 5; LGd: dorsal part of the lateral geniculate complex; TH: thalamus; LD: lateral dorsal nucleus of the thalamus; LP: lateral posterior nucleus of the thalamus; LHA: lateral hypothalamic area; LPO: lateral preoptic area; ZI: zona inerta

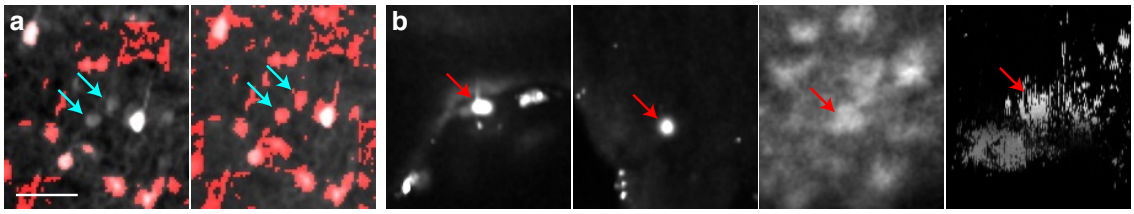


Figure 4-10: Thresholding errors and false positives

a) Example of thresholding errors in tiles containing cells at different brightness levels. Some cells with a low brightness are not accurately thresholded (cyan arrows, left tile). A modification of the thresholding algorithm to perform more detailed analysis of dark areas solves this problem (cyan arrows, right tile). b) Examples of false positives, caused, from left to right, by bright autofluorescent particles (2x), background fluorescence in the cerebellum and bright imaging artifact of unknown origin. All images show STP data from adult Ntsr1-Cre mice injected with rabies virus into the primary visual cortex. Scale bar: 50µm

Detailed analysis of the visual areas revealed, that the disparity between FACCT and human raters was most strongly present in the primary visual area, proximal to the viral injection site. This can in part be attributed to the high density of cells. However, in addition to that, a number of cells in these areas were not detected by FACCT. Further investigation revealed that the threshold algorithm can under some circumstances fail to detect dark cells in the direct vicinity of very bright cells, which appears to be the most likely cause for the majority of the missed cells (Figure 4-10). Interestingly, the number of cells in layer 4 of the primary visual cortex was consistently reported lower than expected, which is likely due to the smaller size of the cells in this layer making them less likely to be detected multiple times, even when there are z-discontinuities in the data.

4.2.8 Z-Discontinuity

All STP datasets examined here exhibited pronounced z-discontinuities (Figure 4-6), leading to cells being detected multiple times. To test how the cell counter performs without these discontinuities, the tile placement was manually corrected in a small region from brain 1. Six human raters were asked to count the cells in a 1400µm*1200 µm *200 µm area to compare the result with the cells marked by FACCT on this dataset.

As expected, the cell numbers reported by human raters and FACCT are similar when the images do not exhibit discontinuities in z (humans: 233, 283, 283, 286, 288, 295; FACCT: 285).

4.3 Discussion

The performance of existing automated cell counting algorithms for 3D whole-brain analysis is difficult to determine as they were either not quantified (Menegas et al., 2015; Vousden et al., 2015) or quantified on a small region in low resolution data (203x203x65 voxels with a voxel size of 4µm (Renier et al., 2016)), potentially limiting the ability of the human rater to accurately identify cells. As a result, the most promising whole-brain cell counting solution has been 2D-based, counting cells on individual images spaced 50µm apart and using a constant correction factor to estimate the true number of cells (Kim et al., 2014b). While this counter also employs deep learning, it uses a relatively simple network architecture with only 3 hidden convolutional layers as opposed to FACCTs ResNet, which uses 9 convolutional layers on its main processing path and 4 to handle the "identity" paths (Figure 2-5). As a result, the cell detection method by Kim et al. was unable to analyse data from the cerebellum and required separate neuronal networks for the olfactory bulb and the rest of the brain, suggesting that it may be challenging to generalise this method for multiple staining or imaging modalities.

An interesting approach based on mean-shift clustering and semantic deconvolution using artificial neuronal networks has recently been described by Silvestri et al. (2015). However, its performance has only been evaluated for a single light-sheet dataset of a mouse cerebellum, for which it employed prior knowledge about the cerebellar anatomy (modelling of cerebellar folds) to reduce the number of false positives. Hence the method's applicability to (noisier) whole-brain STP data remains unknown.

A unique feature of FACCT is its heavy use of data reduction via a cascade of fast image filters. A full-image segmentation approach, as employed by Kim et al. (2014b) would require the deep learning module to analyse ~351,000,000,000 locations in our STP datasets. While it is possible to employ a more aggressive design that classifies a larger space on each run (Silvestri et al., 2015), this would only reduce the number of required runs by 2-3 orders of magnitude. The approach used by FACCT, on the other

hand only requires its deep learning module to check ~60,000 potential locations for a connectivity tracing experiment with ~20,000 labelled cells. As a result FACCT can employ a much more complex artificial neuronal network for cell classification while maintaining a high processing speed. While the filtering and data extraction leading up to the deep learning module take roughly a day, the deep learning module itself can classify 60,000 cell candidates in less than an hour on a 2x6core Xeon workstation with a Nvidia Titan X GPU.

While the ResNet is very capable at removing false positives, its performance is still not perfect, resulting in 0.2-3.6% of detected cells being clear false positives (no human count in a radius of 100µm). This appears to depend mostly on the quality of the dataset (e.g. perfusion quality, particles). Going forward, the high speed and the extendibility of the ResNet architecture would allow to add further hidden layers to reduce the false positive rate, especially when combined with "dropout learning", a method where parts of the network are randomly disabled during training to reduce overfitting (Hinton et al., 2012). Currently FACCT uses 9 hidden layers on the main path, while the original ResNet publication successfully tested up to 1202 hidden layers and reported improvements for increased layer number with up to 110 hidden layers (He et al., 2015a).

While the enhanced Otsu thresholding paradigm generally worked well, there were instances where the algorithm failed to correctly threshold a low-brightness cell in the same tile as a high-brightness cell (Figure 4-10 a, upper panel). This error can lead to under-reporting of cells in areas with dense clusters of cells at different brightness levels, as is the case at the injection site of the RV. A modified version of the thresholder solves this problem (Figure 4-10 b, lower panel) by re-calculating the Otsu threshold in the darker areas of tiles where the first pass of Otsu thresholding resulted in a valid cell structure. However, the overall performance of this modification has not yet been evaluated. If necessary, FACCT's extremely modular design allows for the use of multiple thresholding designs to, for example, employ different algorithms depending on the brightness or number of structures in a particular area.

The biggest remaining issue however is not directly related to the cell detection algorithm, but caused by incorrect stage coordinates from the TissueVision STP microscope, which is exacerbated by the fact that the pipeline assembling the images

attempts to correct the positions of individual tiles without checking for continuity in z. Unfortunately, this introduces significant noise into the data, as whole regions can be shifted by several tens of microns, which leads to repeated detections of the same cell. In dense regions, a single tile shift can lead to over 100 additional cells being detected. Unfortunately, the severity of this issue is not consistent between datasets and even within a dataset there is no apparent way to predict the direction and severity of shifts. Attempts to filter multiple detections of the same cells using their position and image similarity were unsuccessful; hence the most immediate goal will be to implement a stitching pipeline for STP data that does not suffer from discontinuities in z.

Despite these issues, the data on FACCT's cell counting performance is extremely promising. There is a strong correlation between the number of cells found by FACCT and human raters for both, anatomical and cuboid segmentations and its high speed permits further optimisation of the ResNet which could further improve its performance.

Chapter 5. Discussion

The aim of this project was to develop and validate a pipeline for automated detection of fluorescently labelled cells in whole-brain STP scans. To accomplish this goal, two tools were developed and tested.

1. aMAP automates full anatomical segmentation of whole-brain STP datasets. It can segment a brain in 40 minutes at a performance that is comparable to expert human raters (Niedworok et al., 2016).
2. FACCT automates detection of cells in whole-brain STP datasets. It can detect cells in one channel of an STP dataset acquired at 1µm x/y resolution in less than 2 days. While z-discontinuities in the STP data led to overestimation of the cell number, quantification of the result showed very high correlation between the number of cells detected by FACCT and human raters.

5.1 Automation of Brain Segmentation

Any functional mapping experiment requires assignment of anatomical location, as it is vital to know which area of the brain is being investigated. To directly compare experiments both within and across laboratories, it is thus crucial to ensure unbiased accurate and reliable anatomical segmentation of the underlying image data. Classically, this would be done manually by human experts using reference atlases, either in printed or digital form (Paxinos and Franklin, 2004). While this form of segmentation is adequate for small-scale studies or individual tissue sections, it is a subjective and time-consuming task, which makes it infeasible for high-throughput whole-brain imaging studies.

This problem has been addressed in the field of human MRI studies by combining image registration with predefined segmented datasets (brain atlases, (Collins et al., 1995)). The tools and methods developed for human MRI datasets have been used on both STP (Oh et al., 2014) and LSFM data (Menegas et al., 2015), however so far there has been a lack of standardisation and validation regarding their use on these datasets. To address this issue aMAP was developed and released as an open platform (Niedworok et al., 2016). It provides a fully documented, simple, fast and most importantly validated option for anatomical segmentation of STP datasets. Although

quantification of segmentation performance on LSFM datasets is yet to be carried out, the very similar nature of the background fluorescence signal suggests that the results should be comparable.

As automated segmentation depends on anatomical atlases, the availability of high-quality atlas data is crucial. Fortunately, excellent progress has been made in this area over the past years. In the initial phase of aMAP's development, the only atlas available was the original Allen brain Atlas (Lein et al., 2007). This atlas is based on manually segmented serial histology sections that were aligned to an MRI reference brain using only affine registration (Ng et al., 2007). As affine registration, by design, is not able to correct (non-affine) slicing and handling artefacts, the 3D version of the Allen atlas suffered from misalignments between the segmentation and the reference brain in several brain regions, in addition to inconsistencies of anatomical borders along the anterior-posterior axis. In a first step, Kim et al. (2014b) addressed the misalignments by generating their own average brain and manually correcting the anatomical segmentations of the Allen brain atlas. As a result, the fit of the segmentation to the average brain was noticeably improved, while the structure outlines remained comparable to the original Allen brain atlas. For this reason, the Kim et al. (2014b) atlas was used in this thesis to evaluate the performance of aMAP against human raters (who used the original Allen brain atlas for segmentation). However, due to its reliance on the original Allen segmentation, the Kim et al. (2014b) atlas still suffers from inconsistencies along the anterior-posterior axis (Figure 2-1). In addition, the anatomical segmentation underlying both atlases is based on a single specimen (Ng et al., 2007). To address these issues, the Allen institute has released a new common coordinate framework (CCF) on the 27th of October 2016. It is an updated atlas that introduces 3D segmentations based on data from multiple experiments using transgenic reporter mice and rAAV connectivity mapping to determine the location of anatomical structures[5]. While the atlas does not yet contain updated 3D versions of all structures present in the original Allen atlas, it is an important step forward and has hence been used to provide the segmentation for the validation of FACCT.

5 https://www.alleninstitute.org/what-we-do/brain-science/news-press/press-releases/allen-institute-brain-science-announces-mapping-mouse-cortex-3d

However, the continuous improvement of brain atlases can impact comparability between studies because the location of anatomical boundaries in the brain are re-evaluated and partially altered between different atlas versions. This highlights another important aspect of automated segmentation: If segmentation is performed manually, applying an updated atlas to the data would require to redo the manual segmentation on the original images. In contrast, automated segmentation maps every point in the original dataset to its corresponding position within the atlas. This permits publication of the locations of points of interest (e.g. labelled neurons), using the atlas coordinate system. If data is published this way, any updates to the atlas can immediately be applied to it, even after publication, ensuring comparability with newer studies.

5.2 Automation of Cell Counting

Manually counting labelled cells in 3D whole-brains scans is a laborious and time-consuming task that represents a major bottleneck for high-throughput connectivity studies. While many tools have been developed for cell detection in 2D or high resolution 3D datasets (Abbas et al., 2014; Oberlaender et al., 2009; Toyoshima et al., 2016), the large size of STP datasets (~2.5TB per brain) combined with lower resolution (1µm/voxel in x/y, 5µm/voxel in z) and bright cell-like imaging artefacts prevents the use of most pre-existing algorithms.

The adoption of GPUs, originally designed for 3D video games, as general purpose processing devices led to a rapid increase in available processing power in recent years. This has enabled the development of deep convolutional (artificial) neuronal networks that are able to detect objects in images with high accuracy (Krizhevsky et al., 2012; Russakovsky et al., 2015). These ANNs have successfully been used for cell detection in both 2D (Kim et al., 2014b) and 3D datasets (Silvestri et al., 2015), making them promising candidates to automate cell-detection in whole-brain STP datasets. However, as of today the large size of a whole-brain STP scan prohibits the use of more complex ANN architectures on full datasets. To overcome this, Kim et al. (2014b) used a simple ANN architecture with 3 hidden layers and only analysed 2D images spaced 50µm apart, whereas Silvestri et al. (2015) used a network with 2 hidden layers to enhance the signal in images of the cerebellum (“semantic deconvolution”), before detecting labelled cells using a classic mean-shift approach that detects clusters of bright signal.

As it is time-consuming to train and apply ANNs to very large datasets, I take the opposite approach to Silvestri et al. (2015). Images are pre-processed using a series of fast filters that detect the location of potential cells, which are then classified using an ANN. On the data presented here, pre-processing took less than two days and reduced the amount of data that needed to be processed by the ANN by at least 2 orders of magnitude. This data reduction enabled the use of a more complex ResNet with 9 hidden layers and allows for further extension if required.

The results presented here demonstrate a strong correlation between manual and automated counts, resulting in a correlation of $r > 0.99$ with the exception of a single dataset ($r = 0.98$) that contained shadow artefacts in parts of the brain. Unfortunately, the presence of strong z-discontinuities in the STP data resulted in multiple detections of individual cells and prevented a conclusive analysis of counting performance in a whole-brain dataset. However, FACCTs result was in line with that of human raters on a small subvolume with manually corrected z-discontinuities.

Going forward, the first step will hence be to implement a stitching pipeline for STP data that corrects misplaced tiles while ensuring consistency in tile placement along the anterior-posterior axis. This will allow a conclusive cell-by-cell comparison between FACCT and human raters.

Next, it would be desirable to quantify FACCT's performance on LSFM and confocal microscopy data. The centroid-detection paradigm of FACCT's initial filters should be applicable to other fluorescence imaging methods. In combination with the general capability of ANNs to detect a multitude of objects using a single network, it should be possible to accurately detect cells in multiple imaging modalities.

The final aim is to release FACCT to the community as an open tool, complete with a pre-trained ANN to enable the immediate cell detection on a variety of 3D fluorescence datasets.

5.3 Applications

Generally, the use of transsynaptic RV tracing in combination with mouse lines expressing Cre recombinase permits labelling of monosynaptic input to genetically defined cellular types. This presents the unique opportunity to characterise neuronal connectivity at the level of individual microcircuits and has already been used to improve our understanding of the connectivity and function of a variety of brain areas such as the cortex (Fu et al., 2014), or midbrain areas (Lammel et al., 2012). However, in an analogue to how automated fast high-throughput sequencing has revolutionised the field of genetics (1000 Genomes Project Consortium, 2012), automated high-throughput analysis of whole-brain connectivity has the potential to greatly advance our understanding of the function of the brain.

Automated high-throughput connectivity analysis will help uncover the changes in brain wiring in mouse models of neuropathological diseases. Such models exist for a variety of conditions such, as Down's syndrome (Li et al., 2007), autism (Peça et al., 2011) and schizophrenia (Sigurdsson et al., 2010), that affect the function of the brain, but the changes in connectivity in these models remain largely unknown. Human functional connectivity data acquired using resting state fMRI (rfMRI) is available (Broyd et al., 2009) and could be used to identify potential targets for further investigation in the respective mouse models. Furthermore, recent advances in rodent MRI methods have enabled mouse rfMRI at high resolution (Desai et al., 2011). Combining RV tracing with rfMRI in mouse models of disease would allow to directly compare the changes seen in human rfMRI with those in the mouse model and analyse the anatomical connectivity changes in the same animal using RV combined STP. This could provide three key insights: Firstly, it would advance our understanding of the relationship between functional connectivity, as reported by rfMRI, and anatomical connectivity. Secondly, it could give valuable information on how well changes in functional connectivity in human patients are replicated in the mouse model, and finally it could provide insight into how anatomical connectivity is changed in neuropathological diseases.

On a more fundamental level, the possibility to rapidly map connectivity in large numbers of animals would also allow us to investigate the role and extent of subject-to-subject variability in neuronal connectivity.

5.4 Outlook

The last few years have seen a rapid growth in the field of automated whole-brain scanning. A new open STP design has been developed and released (Economo et al., 2016), and the introduction of a multitude of new clearing methods (Chung et al., 2013; Lee et al., 2016; Renier et al., 2016; Schwarz et al., 2015) are likely to lead to more widespread adoption of LSFM in the rodent neuroscience community. This process could result in a marked shift, enabling small research groups to carry out high-throughput studies that have previously been reserved to dedicated highly funded research centres such as the Allen Brain Institute.

To facilitate this process, the development of analysis software tailored to large 3D datasets is of utmost importance. While this thesis introduces tools for automated analysis of STP data that can hopefully be adapted to other whole-brain imaging modalities, handling and visualisation of these datasets remains a challenge. Due to their size, it is not possible to fully load a whole-brain STP scan into memory, and opening a partial dataset can take tens of minutes. In line with that, the processing time of FACCT is dominated by loading and saving of files. Iterating over the images of a single channel without performing any other operations currently requires ~3 hours of processing time, even when the data is loaded from a fast solid-state disk. This is exacerbated by the fact that our current data structure (one tiff file per coronal image) does not permit partial loading within an individual image. As a result, it is necessary to load a complete coronal image, even if only a small fraction of that image needs to be analysed.

Currently, there are two possible candidates to solve this problem. BigDataViewer is an extension to ImageJ that stores images in the HDF5 file format and enables quick 2D visualisation at arbitrary viewing angles (Pietzsch et al., 2015). It stores images at multiple resolutions and allows to quickly open and view terabyte-sized datasets on regular desktop computers. The downside is that it requires data to be converted to

HDF5, and manipulation of the files after conversion is difficult, as ImageJ does not provide direct write access to the format. As a result it currently offers only basic viewing functionality.

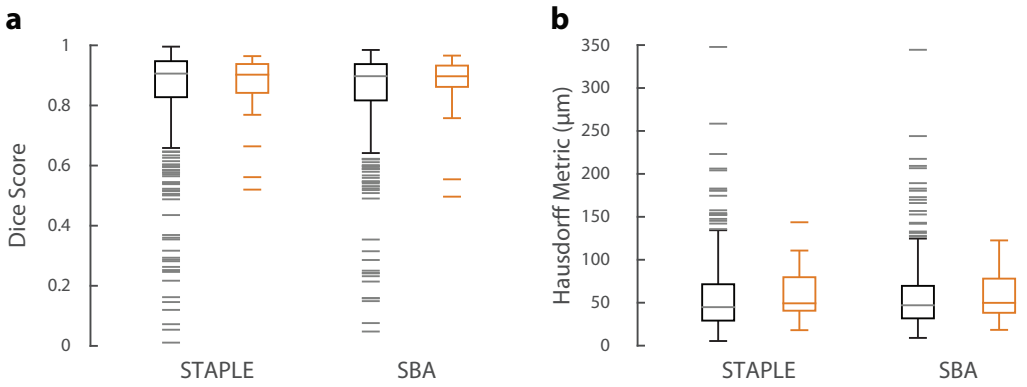
The other candidate is Vaa3D, an open source toolkit to view and visualise microscopy data (Peng et al., 2014). Vaa3D is maintained by Janelia Research Campus and the Allen Brain Institute and seems to be aimed specifically at large datasets. It is under active development and has recently been extended to allow stitching and viewing of terabyte-sized datasets using a custom format based on tiled tiff images at multiple resolutions (Peng et al., 2015). While we have experienced a number of stability issues with the software, it has made huge process in the past years and is likely to soon become a valuable platform for visualisation and analysis of large 3D datasets.

In the field of viral tracers, research on different RV strains and glycoproteins have led to more efficient and less neurotoxic variants of the virus that label a higher proportion of the presynaptic input and allow functional investigation of the connected network by expression of optogenetic tools or calcium indicators (Kim et al., 2016; Reardon et al., 2016; Yamawaki et al., 2016).

In summary, recent advances in automated microscopy, combined with powerful genetic tools have given us the ability to map the connectome with unprecedented speed and detail. The software developed for this thesis aims to help in that endeavour by automating the most laborious and time-consuming parts of the analysis. Given the recent advances in the technology, I am confident that we will see a broader adoption of high-throughput connectivity analysis. The result will hopefully be a critical mass of connectivity data that, combined with further functional investigation, would greatly advance our understanding of the brain.

Chapter 6. Appendix

6.1 Additional Accuracy Measures



Box plot showing the Dice scores of manual raters (n = 22 raters, each segmenting 2 out of 4 potential brains, grey) and aMAP (orange) for nine structures in 4 brains, based on the consensus-segmentation generated by STAPLE and SBA, respectively. There were no significant differences in median scores of human raters versus aMAP. (STAPLE: 0.91 vs. 0.90, p=0.5; SBA: 0.89 vs. 0.90, p=0.8; Mann-Whitney-U Test) **b)** Box plot showing the pooled Hausdorff metrics of manual raters (grey) and the aMAP segmentations (orange) on the same brains and structures used in a. There were no significant differences in median scores of human raters and aMAP. (STAPLE: 45µm vs. 49µm, p=0.06; SBA: 47µm vs. 50µm p=0.27; Mann-Whitney-U Test).

6.2 Exemplary Segmentations of the Dentate Gyrus, Granule Cell Layer (DG-sg)

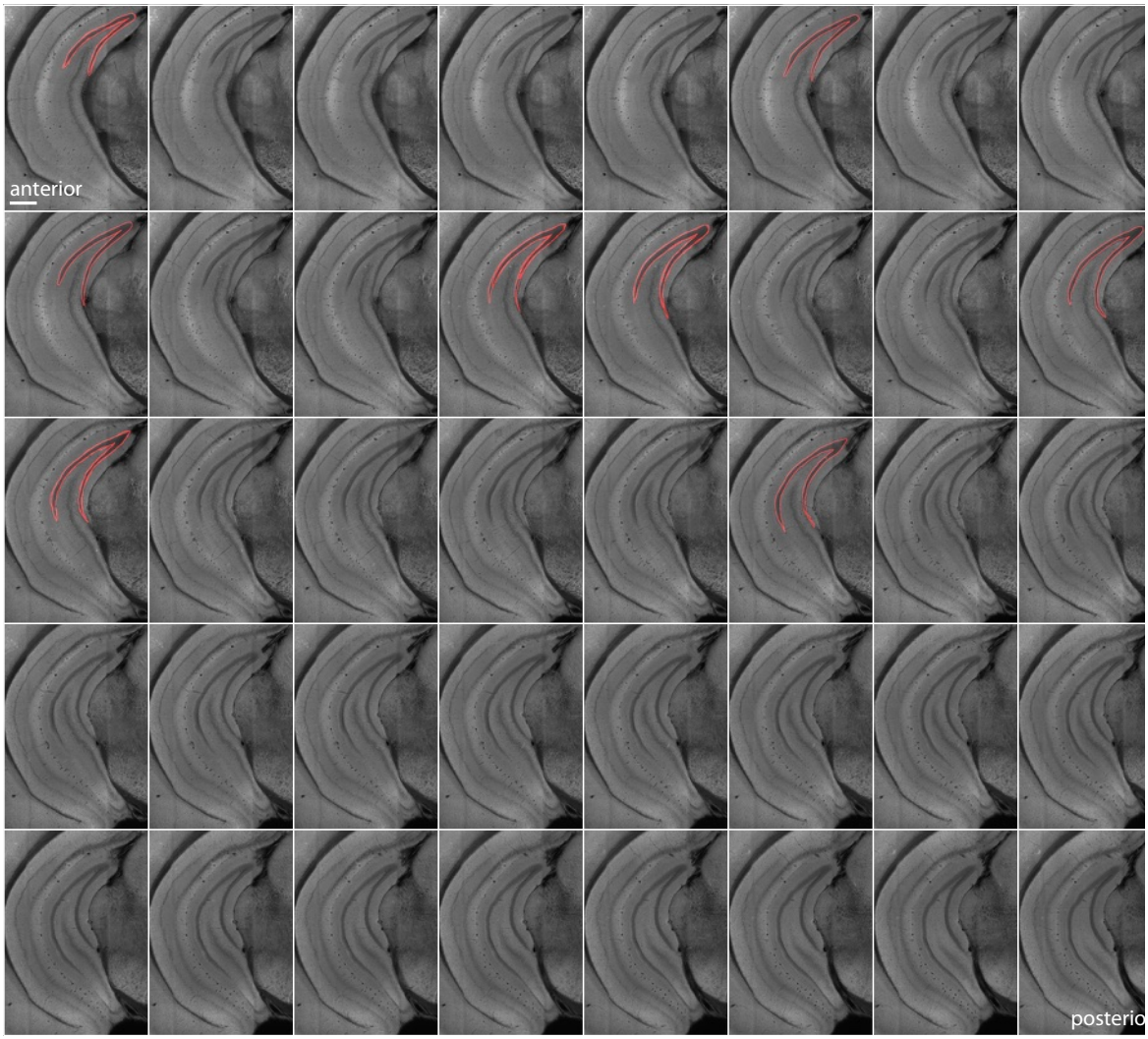


Image series showing the area around the hippocampus from the DG-sg segmentation task in one brain, ordered from anterior to posterior. Raters' outlines are highlighted in red. Note the distinct change in the shape of the clearly identifiable DG-sg outline from anterior to posterior. Scale bar = 500µm.

nature COMMUNICATIONS

ARTICLE

Received 10 Feb 2016 | Accepted 9 May 2016 | Published 7 Jul 2016

DOI: 10.1038/ncomms11879 OPEN

aMAP is a validated pipeline for registration and segmentation of high-resolution mouse brain data

Christian J. Niedworok[1,2], Alexander P.Y Brown[1,2], M. Jorge Cardoso[3], Pavel Osten[4], Sebastien Ourselin[3], Marc Modat[3] & Troy W. Margrie[1,2]

The validation of automated image registration and segmentation is crucial for accurate and reliable mapping of brain connectivity and function in three-dimensional (3D) data sets. While validation standards are necessarily high and routinely met in the clinical arena, they have to date been lacking for high-resolution microscopy data sets obtained from the rodent brain. Here we present a tool for optimized automated mouse atlas propagation (aMAP) based on clinical registration software (NiftyReg) for anatomical segmentation of high-resolution 3D fluorescence images of the adult mouse brain. We empirically evaluate aMAP as a method for registration and subsequent segmentation by validating it against the performance of expert human raters. This study therefore establishes a benchmark standard for mapping the molecular function and cellular connectivity of the rodent brain.

[1] The Division of Neurophysiology, MRC National Institute for Medical Research, London NW7 1AA, UK. [2] The Sainsbury Wellcome Centre for Neural Circuits and Behaviour, University College London, London W1T 4JG, UK. [3] Translational Imaging Group, Centre for Medical Image Computing, University College London, London WC1E 6BT, UK. [4] Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. Correspondence and requests for materials should be addressed to M.M. (email: m.modat@ucl.ac.uk) or to T.W.M. (email: t.margrie@ucl.ac.uk).

Brain-wide mapping of neuronal gene expression[1], connectivity[2–4] and function[5] is required if we are to obtain a complete understanding of the physiological processes underlying cognition and behaviour. Recent advances in tissue clearing and high-resolution light microscopy[6–11] combined with modern transgenic and neuronal tracing methods[12,13] now make mapping of the mammalian brain with cellular resolution a feasible prospect[1–3,14–17]. However, any mapping effort requires the implementation of an objective, accurate and reliable means of defining the anatomical boundaries of underlying brain structures. The accuracy of this segmentation process is dependent on image registration and is critical since it defines the identity of cells or neuronal connections in terms of their anatomical position, a process that underpins interpretation and comparison across experiments.

Recently, automated high-resolution microscopy instruments have dramatically increased the throughput of data acquisition[8,11] rendering manual segmentation an unfeasible prospect and necessitating the development of automated analytical pipelines. The most common approach for automating anatomical segmentation is called atlas propagation and involves performing registration of an image data set onto a standardized, fully segmented reference space to provide an anatomical segmentation of the original images[1,2,14]. One critical aspect regarding the implementation of such pipelines is ensuring that the quality of the resulting segmentation—previously achieved by expert neuroanatomists relying on their experience and detailed visual inspection of the data—is not compromised.

Such high-throughput microscopy instrumentation produces large volumes of high-resolution three-dimensional (3D) data and relies on the accuracy of automated segmentation, yet to date there has been only indirect assessment of segmentation quality and no agreement on a standard method of implementation, with individual labs using unpublished in-house tools[2,14,16] or an open source clinical image registration tool (Elastix[18]) with unpublished parameters[1,17]. While these tools may perform adequately in their respective labs, only a validated, open source and fully automated method can enable the direct comparison of emerging data sets and cross-laboratory agreement.

Here we present aMAP, a tool that internally uses and provides a graphical front-end to NiftyReg (a rapid image registration toolkit, originally developed for human MRI data[19]), that we modified to enable rapid processing of high-resolution 3D light microscopy data. aMAP permits propagation of a 3D mouse atlas of the entire adult mouse brain in 40 min and its accuracy and reliability is shown to be on par with expert human raters.

Results

Assessing segmentation quality. To assess the performance of human raters on manual segmentation, twenty-two neuroscientists were randomly assigned to one of two groups and asked to segment the same ten target structures (of which 9 were analysed; see Methods; Fig. 1a) from three brain datasets. Target structures were presented within six serial two-photon (STP) image stacks (40 coronal planes per stack containing tissue background fluorescence ($n=5$ brains) or sparse red fluorescent protein (RFP) labelling ($n=1$ brain)) obtained from adult C57BL/6 mice. These structures were chosen to encompass a broad range of sizes and anticipated difficulty based on their degree of border definition according to local anatomical landmarks. Raters were required to identify one image plane from the STP stack that best matched the target section presented from the two-dimensional (2D) anatomical reference atlas of the Allen Brain Institute[14] and then asked to manually outline the perimeter of the target structure on the STP image (see Methods).

Qualitatively, human raters showed substantial inter-rater variability in their positioning of the borders and estimation of the size of target structures (Fig. 1b). In general, there was stronger agreement—that is low inter-rater variability—where the structure could be identified using high-contrast landmarks, such as structure borders at the ventral and dorsal surfaces of the neocortex. However, we observed weak agreement (high inter-rater variability) at borders that were less well delineated, such as the cortico-cortical boundaries of cortical target regions (for example, primary visual cortex (VISp), Fig. 1b). Disagreement between raters was particularly significant for target structures that lacked any distinct anatomical landmarks, such as the ventral posteromedial nucleus of the thalamus (VPM), the segmentations of which, showed very little overlap in boundary definition (Fig. 1b).

Recent approaches to validating mouse-brain segmentation have relied on comparing the Euclidean distance between manually chosen anatomical landmarks in an image data set before and after its registration (step 1; Supplementary Fig. 1) to an average brain image[1,9]. While this method is easily implemented, it can only report registration accuracy proximal to the chosen landmarks and is not indicative of the quality of segmentation (step 2; Supplementary Fig. 1).

On the other hand, direct assessment of segmentation quality is hindered by the fact that there is no 'ground truth' regarding the precise location of an anatomical structure in any data set. Thus, it is not possible to assess the quality of either automated or human segmentations without first establishing a 'ground truth' segmentation of the image data sets. To achieve this essential initial step we therefore determined the consensus segmentation of all human raters for each target structure in each brain data set using STAPLE[20], an iterative algorithm that—when given multiple segmentations—simultaneously estimates the quality of each segmentation and derives the quality-weighted consensus (see Methods). Using the STAPLE consensus segmentation as a 'ground truth' we could now directly evaluate segmentation performance of both human raters and aMAP using the Dice score metric[21] that quantifies the overlap between two structures and is commonly used to assess automated segmentation quality[22].

Consistent with the idea that the STAPLE–Dice method is directly reporting the quality of segmentation, we first determined that imposing a goodness of fit on the registration of STP images to an average brain data set[1,9] (that is, by constraining the bending energy) exerted a significant influence on Dice scores that improved with increasing bending energy weight (repeated measures ANOVA, $F_{(15,75)}=16.8$, $P<0{,}001$; Supplementary Fig. 2a (range 0.2–0.95)). In contrast, the Euclidian distance between landmarks was insensitive to changes in the goodness of fit of the registration imposed by the same range of bending energy weight (repeated ANOVA, $F_{(15,75)}=0.45$, $P=0.95$ Supplementary Fig. 2b).

To score segmentation quality of both human raters and aMAP, we next compared each segmentation with the appropriate STAPLE consensus using the Dice score. As a complimentary measure, we also used shape-based averaging (SBA)[23] to generate an average segmentation of human raters and the Hausdorff metric as a second segmentation quality metric (see Methods, Supplementary Fig. 3a,b). Although these different methods for determining the ground truth segmentation and segmentation quality produced very similar results (Supplementary Fig. 3a,b), we adopted the STAPLE–Dice metric, as it is the most widely accepted analytical tool used in other imaging fields[22].

aMAP was implemented using the open-source NiftyReg toolkit[19] to register the average brain of the Kim *et al.*[1] 3D atlas to downsampled versions of our STP data sets (12.5 μm

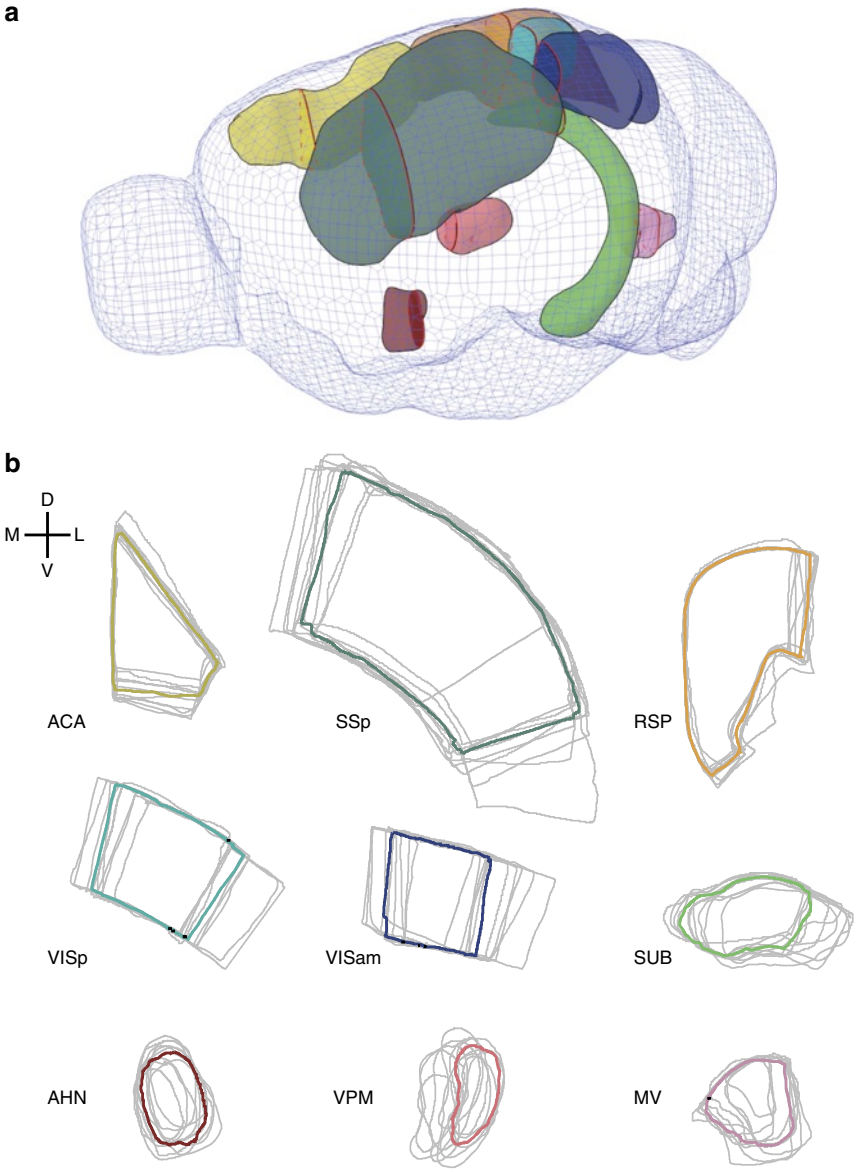


Figure 1 | Anatomical structures used to assess segmentation performance. (**a**) An illustration showing the 3D shape of the nine brain structures in the left hemisphere used to assess segmentation performance. Red lines within each structure highlight the coronal plane in the reference atlas that was presented to human raters. (**b**) For each structure, the segmentation outlines are shown for a given group of 11 raters (grey lines). The consensus outline for the same structure and 11 raters as determined by STAPLE is overlaid (bold coloured line). According to The Allen Brain Atlas nomenclature, the nine structures shown are: anterior cingulate area (ACA); anterior hypothalamic nucleus (AHN); medial vestibular nucleus (MV); retrosplenial cortex (RSP); primary somatosensory area (SSp); subiculum (SUB); primary visual cortex (VISp), secondary visual cortex, anteriomedial part (VISam); ventral posteromedial nucleus of the thalamus (VPM) (D: dorsal; V: ventral; M: medial; L: lateral).

isotropic) using affine and free-form registration. The resulting transformations were then applied to the 3D Kim *et al.*[1] brain atlas, which is based on the Allen Institute Brain Atlas that was used here by human raters. Computation time for aMAP-based segmentation was 40 min per downscaled brain on a dual-6-core Xeon workstation. To find the appropriate parameters for the image registration, we used two of the six STP brains as training sets. Unless noted, these brains were excluded from the analysis of aMAP's performance.

Human raters versus aMAP. The outlines obtained from aMAP (Fig. 2a, orange lines) were qualitatively similar to those performed by human raters (Fig. 2a, grey lines), which was confirmed by Dice score analysis (Fig. 2b–d). When pooling the scores of all structures, the median score achieved by aMAP was not significantly different from human performance levels (Mann–Whitney *U*-test, score of 0.92 versus 0.91, $P=0.52$; $n=4$ brains, 9 structures, 22 human raters). When grouping these scores by structure, there were no significant differences between the scores for human raters and aMAP in eight out of nine structures. Humans scored significantly better in segmenting the anterior cingulate area (ACA, Mann–Whitney *U*-test, median Dice score of 0.952 versus 0.870, $P=0.005$; Fig. 2b). When grouping the scores by brain rather than structure, there were no significant differences observed between human raters and aMAP for any individual brain (Mann–Whitney *U*-test, $P>0.49$, Fig. 2c).

However, despite there being no significant difference in the overall median scores between human raters and aMAP, human raters exhibited substantial variance, while the Dice scores

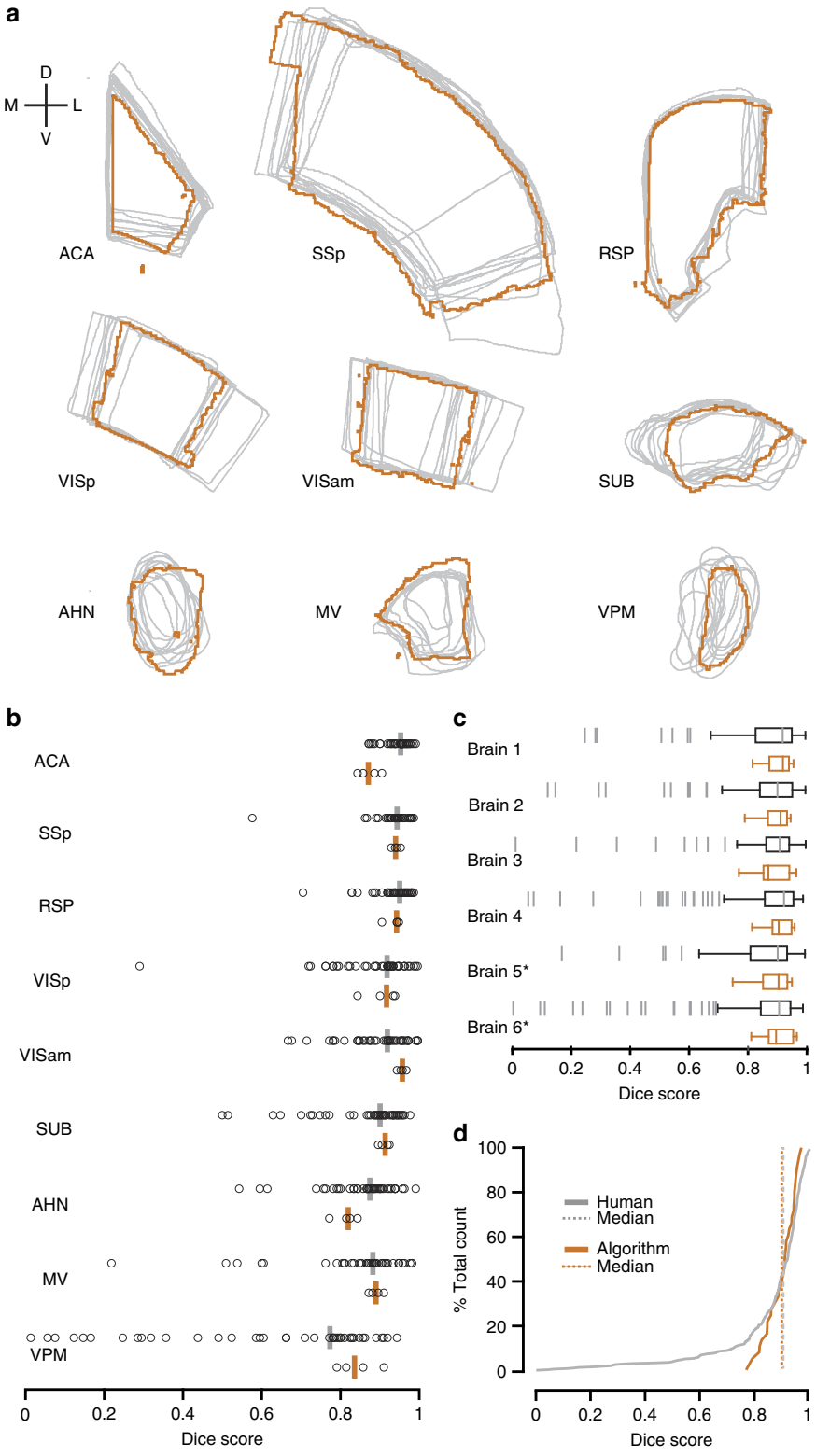


Figure 2 | Segmentation performance of human raters and aMAP. (**a**) Segmentation outlines of human raters (grey) with the aMAP segmentation result of the same structure and brain overlaid (orange). (**b**) Dice scores for manual ($n=22$ raters, each segmenting two of four potential brains, grey) versus aMAP ($n=4$ brains, orange) segmentations grouped by target structures ($n=9$). (**c**) Box plots showing Dice scores of human (grey) versus aMAP (orange) segmentations grouped by brain. Brains used in the registration parameter search (training data) are marked with an asterisk. (**d**) Cumulative histogram of the Dice scores for manual (grey) and aMAP (orange) segmentations for all structures and brains as shown in **b**. Vertical lines indicate the median scores.

obtained by aMAP were significantly more consistent (Levene's test on pooled scores; $n=4$ brains, 9 structures, 22 human raters; s.d.: 0.16 versus 0.05, $P=0.005$, Fig. 2d). In addition to the variance observed for x–y border definitions, human raters strongly disagreed as to which optical section of the STP data sets best corresponded to the Allen Brain Atlas plates, leading to substantial variation in the identity of the section chosen for segmentation (z-choice, Fig. 3a; median anterior–posterior distance between two raters on the same brain and structure ($n=6$ brains)—ACA: 15 µm, primary somatosensory area (SSp): 120 µm, retrosplenial cortex (RSP): 45 µm, VISp: 75 µm, secondary visual cortex, anteriomedial part (VISam): 60 µm, subiculum (SUB): 75 µm, anterior hypothalamic nucleus (AHN): 105 µm, medial vestibular nucleus (MV): 112.5 µm, ventral posteromedial nucleus of the thalamus (VPM): 105 µm, dentate gyrus, granular cell layer (DG-sg): 90 µm). Such differences in z-choice had a particularly strong influence on segmentations of DG-sg that resulted in substantial discrepancies in x–y border definitions (Fig. 3b), despite the fact that the structure could be clearly delineated in the STP data set (Supplementary Fig. 4). Thus while manual segmentations performed on the same optical plane of the dentate gyrus generally showed good agreement (Fig. 3c), the overall x–y boundary of the DG-sg changed substantially according to the choice of z-section (Supplementary Fig. 4). We therefore excluded this structure from the segmentation analysis, since the discrepancies in segmentations were substantially negatively influenced by differences in z-choice rather than a rater's uncertainty about the x–y boundary of the structure. We found no significant influence of the z-choice range on the median Dice scores of human raters for the remaining structures ($P>0.05$, Supplementary Fig. 5).

Intra-rater reliability. By design, aMAP will always produce an identical segmentation when applied to the same data set. In contrast, one major source of the significant inter-rater disagreement on manual segmentations could be a rater's degree of reliability. Although an individual may score poorly compared with the STAPLE consensus, they may nevertheless be extremely reliable in their estimate of the location and shape of the target structure (Fig. 3d). On the other hand, the trial-to-trial reliability of a rater could significantly contribute to the broad range of (inter-rater) Dice scores. Reliability could therefore be considered to be one major source of variability inherent in the segmentation process and distinct from inter-rater disagreement.

To investigate the extent to which inter-rater disagreement in the segmentation stemmed from an individual's uncertainty or

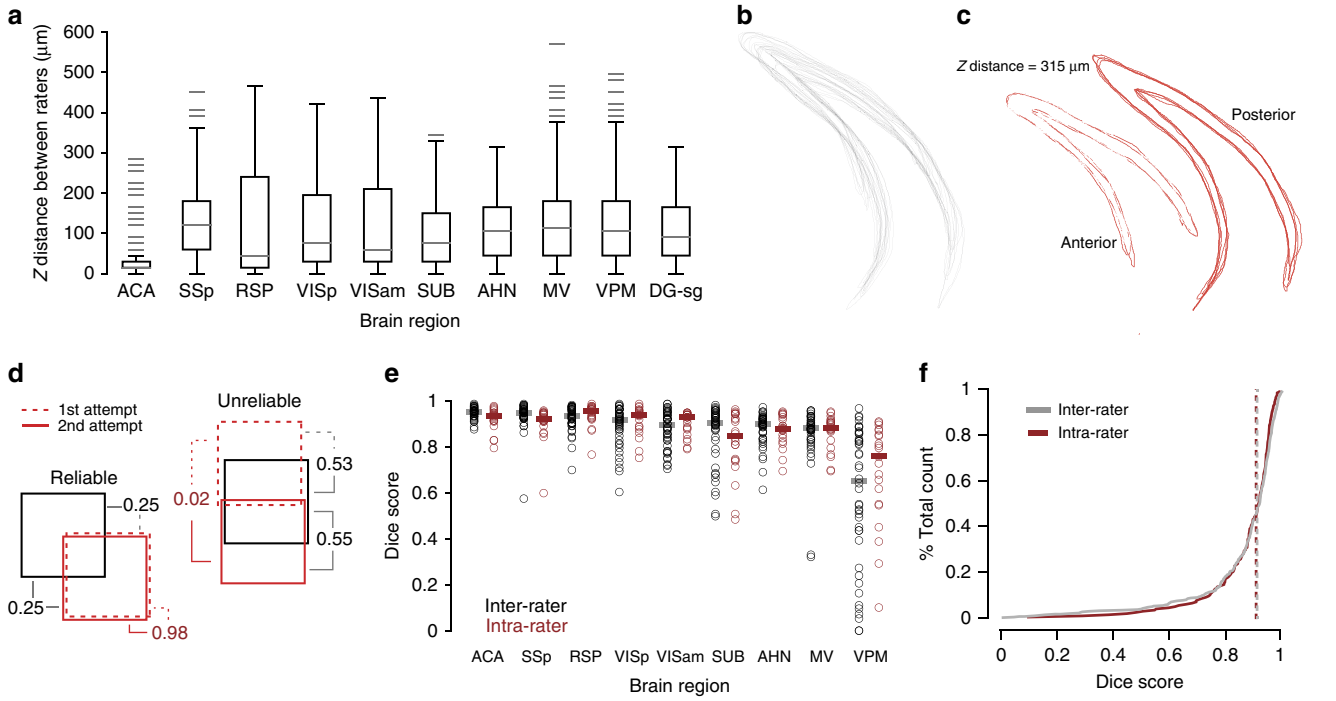


Figure 3 | Sources of variance in manual segmentation. (**a**) A box plot showing the anterior–posterior distance between any two human raters ($n=22$ raters, each segmenting three of six potential brains) in their estimation of the correct optical section (z-choice) for manual segmentation for each brain structure. Structures: anterior cingulate area (ACA); anterior hypthalamic nucleus (AHN); dentate gyrus, granule cell layer (DG-sg); medial vestibular nucleus (MV); retrosplenial cortex (RSP); primary somatosensory area (SSp); subiculum (SUB); primary visual cortex (VISp); secondary visual cortex, anteriomedial part (VISam); ventral posteromedial nucleus of the thalamus (VPM). (**b**) Example manual segmentations ($n=22$) of the DG-sg performed by 11 human raters taken from a single test brain and its repeated presentation. (**c**) Example manual segmentations of the DG-sg, taken from two z-sections from within the data set shown in **b**. These two z-sections were chosen based on their having multiple segmentation attempts ($n=4$ outlines shown in each image, left image: anterior, right image: posterior). (**d**) Schematic highlighting two extreme segmentation reliability scenarios. Bottom left: a given rater may perform poorly against the STAPLE consensus (black square) of all raters (Dice score = 0.25, grey lines). However, generation of a Dice score that determines the overlap between the first attempt and the second attempt (intra-rater Dice score, red) indicates high reliability (for example, Dice score = 0.98). In contrast, a given rater may obtain a Dice score more similar to the STAPLE consensus but be unreliable in their estimate of the location of the structure (for example, intra-rater Dice score = 0.02; top right). (**e**) Plot of the inter-rater ($n=44$ segmentations per structure; that is, first and second attempt versus STAPLE consensus for 22 raters per structure, black) and intra-rater ($n=22$ segmentations, that is, second attempt versus first attempt for 22 raters per structure) Dice scores for each target structure. (**f**) Plot showing the cumulative histogram of intra- and inter-rater Dice scores for all data presented in **e**.

from a reliable difference in opinion between raters, each user was unknowingly also presented with repeats of all target structures from one of their three previously presented brains. This permitted calculation of a Dice score between a rater's first segmentation and the repeat segmentation of the same structure in the same data set (intra-rater Dice, Fig. 3d). Comparison of the intra-rater Dice score with the previously determined inter-rater scores from the same brains and structures (Fig. 3d) showed significantly worse intra-rater performance on the ACA and SSp, (Mann–Whitney *U*-test, inter versus intra: ACA: median 0.953 versus 0.933, $P=0.01$; SSp: median 0.950 versus 0.923, $P=0.001$; Fig. 3e, $n=2$ brains $\times$ 9 structures $\times$ 11 raters per brain) and no significant difference on the remaining structures (Mann–Whitney *U*-test, $P>0.21$). There was no significant difference in the overall median of inter- versus intra-rater scores (Mann–Whitney *U*-test, inter versus intra: 0.916 versus 0.912; $P=0.32$) and only a modest but significant reduction in variance (Levene's Test, inter versus intra: s.d.; 0.16 versus 0.12, $P=0.044$, Fig. 3f). This indicates that for a given rater, there exists substantial variability in repeated segmentation of the same structure. Thus, human inconsistency is a significant source underlying segmentation disagreement between individual raters.

Discussion

Any attempt to map the brain with cellular resolution depends critically on an objective, accurate and reliable means of defining its underlying architecture. In this study, we have directly compared the performance of an algorithm for automated segmentation of high-resolution 3D fluorescence data sets with the segmentation performance of human raters. We show that manual segmentation is a process that, on average, is of high quality but with modest reliability. While human raters—as a group—generally achieved high median scores, they displayed significant variability, particularly for structures that did not follow obvious anatomical landmarks. The fact that this variability was at least as large for intra-rater comparisons as it was between raters highlights that both the accuracy and reproducibility of manual segmentation is inherently limited. On the other hand, aMAP performed just as accurately as human raters, but with significantly less variability and, by design, is entirely reproducible.

We have designed aMAP based on NiftyReg because of the high speed of its free-form registration and the possibility to adapt it to large data sets. There are however several other applicable MRI registration tools in use in the clinical field[24] that may be equally suitable for 3D fluorescence data. Previous rodent brain microscopy studies have used pipelines based on such tools (for example Elastix[1,17] and MNI AutoReg[16]) or unpublished in-house tools[2,14], but did either not publish validation data of their image analysis pipeline[16,17], or validated their segmentation by relying on a landmark distance-based measure that determines the Euclidian distance between a limited number of point markers in the registered data set[1,2,14]. Here we show that the landmark distance metric does not capture changes in registration quality over a wide range of deformations imposed on the image data set that, by its very nature, impacts the quality of the segmentation process. It is worth noting, that results from previous studies relying on the distance between point markers may nevertheless be accurate. However, our data mirrors previous findings showing that such scoring metrics do not capture the quality of free-form registration of MRI data sets[25]. To encourage community-wide implementation and validation of automated segmentation tools, we have made our manual segmentation data and validation pipeline freely available (see Methods).

In our study, we have ensured parallelity of the optical sections to the coronal plane of the atlas by rigidly registering all images to the Allen average brain. Furthermore, we specified to the human rater the atlas sections that contain the target structure. Nevertheless, differences in the *z*-plane chosen by the raters from the optical stack could remain a significant source of inter-rater variability in segmentation performance. However, at least for the structures analysed here, we found that Dice scores were not significantly improved when our segmentation analysis was confined to optical planes within seven or three sections of one another. It is of course also conceivable that in the real-world scenario both image misalignment and lack of agreement on the correct atlas section could further increase inter-rater variability.

Agreement can be achieved by using several experts to cross-validate segmentations, a practice that is widely used on MRI data in the clinic[20]. However, since high-resolution whole-brain fluorescence data sets are typically several orders of magnitude larger than MRI data sets, this approach is extremely difficult to implement without substantially down-sampling and thereby compromising accuracy. Also, manually agreeing on brain-wide segmentation of high-resolution images is a very laborious and time-consuming process. Particularly for high-throughput pipelines, validation using a limited number of agreed expert raters is impractical and would slow what is already a major analytical bottleneck.

The success of registration depends on the similarity of the images being registered to one another. As such, the location and integrity of key anatomical landmarks (such as the cortical surface) are critical to accurate brain registration. To maximize similarity between our data and the atlas average brain data set (which was generated using tissue autofluorescence) we have used either the background fluorescence or a sparsely labelled RFP channel. In contrast to using fluorescence images exhibiting for example, a very specific anatomical pattern of GFP, this ensures that most pixel values in the image reflect anatomical structures. While aMAP can theoretically be used on image data containing fluorescent signals, it is not possible to reliably predict the impact of such signal patterns on the registration process. We therefore recommend manual quality assessment of the images and their segmentation, especially in cases where specimens have suffered dissection-related damage or that contain excessive imaging artefacts, such as high non-specific background fluorescence (e.g. due to a failed perfusion). We found that overlaying the original image data with the registered average brain and the segmentation outlines provides a reasonable way to qualitatively assess image registration and segmentation.

One shortcoming of all current atlas-based automated segmentation approaches arises from the fact that existing 3D atlases either have (i) adequate 3D segmentation but contain a limited number of annotated structures, as is the case for mouse MRI atlases[26–28] or (ii) have a reasonable number of annotated structures but are based on reconstructions of serial 2D sections rather than genuine 3D segmentations[1,14]. This latter scenario unfortunately leads to discontinuity in structure borders in the plane orthogonal to the atlas's cutting plane[29] that will propagate into any automated segmentation based on such an atlas (Fig. 2a; Supplementary Fig. 6). Despite this limitation, aMAP nevertheless performs on par with human raters and its implementation provides a means of establishing an agreed standard for automated segmentation. Fortunately, in its most recent release, the Allen Brain Institute has begun to move towards a higher resolution 3D-segmented atlas. Although this recent version contains a mixture of 2D and 3D annotations, the goal is to eventually generate an atlas that is fully annotated in 3D. This represents a crucial step forward that will further improve the quality of automated segmentation.

A recent development in the field of MRI imaging has been the introduction of multi-atlas registration to increase the robustness

of automated segmentation[30–33]. There, multiple atlases are registered to the data set of interest and the final segmentation is generated using the consensus segmentations from all individual atlases. While this method increases the robustness of the automated segmentation, the high-resolution multi-atlas datasets necessary to implement this method on 3D light microscopy data do not yet exist.

The fact that automated segmentation will, by design, adapt to future atlas releases highlights another important aspect: Automated segmentation generally works by mapping all points of interest in the experimental data (for example, neuronal and glial somata) to a common reference space. As refinements are made in the segmentation of this common space or new areas are functionally delineated, tools such as aMAP can be used to systematically apply these changes to existing and previously published data sets (assuming points of interest are published using reference space coordinates). In this way, pipelines such as aMAP will enable the data of previous and future studies to be directly compared as 3D mouse atlases evolve.

In summary, we have for the first time validated a tool for segmenting high-resolution 3D imaging data that will rapidly register and segment a complete adult mouse brain data set. aMAP performs as well as human raters but with substantially less variability and thus enables direct comparison of anatomical data sets independent of the level of experience and knowledge base of the user. aMAP can therefore be used to standardize the segmentation process and enable comparability of data from one individual and lab to another. Furthermore aMAP will, by design, inherently adapt to any future refinements in digital segmentation atlases, the precise application of which is currently a significant factor limiting the accuracy of brain-wide mapping approaches.

Methods

Imaging. Male adult C57BL/6 mice were trans-cardially perfused with cold 4% PFA-solution under general anaesthesia. Brains were then removed and post-fixed in 4% PFA for at least 24 h. All procedures were in accordance with UK Home Office regulations (Animal Welfare Act 2006) and the local animal ethics committee. Brains were imaged coronally at a voxel size of 1 µm (x) × 1 µm (y) × 5 µm (z) under a STP microscope[9] using an Olympus ×10 water immersion objective (numerical aperture 0.6). The STP image files for all target brains were rigidly aligned to the 3D average brain of the Allen Mouse Brain Atlas[14] to ensure optical sectioning in the coronal plane of all datasets. The transformation matrices were determined on z-smoothed (Gaussian, 5 voxel s.d.) and then downscaled versions of background fluorescence ($n = 5$) or sparsely labelled RFP ($n = 1$) images using NiftyReg (reg_aladin[34], voxel size 12.5 µm isotropic, https://sourceforge.net/projects/niftyreg). The resulting transformation matrices were then applied to the full-resolution images using MATLAB (MathWorks).

Segmentation task. Manual segmentation data was obtained from a group of 22 neuroscientists that included postgraduate students, research assistants, postdoctoral fellows and principal investigators. This cohort was randomly split into two groups ($n = 11$ raters per group) whereby every rater from within a group performed 10 segmentations on each of three brains (three different brains per group). In addition, all 10 structures from one of the three brains were re-presented blindly as a fourth data set to assess intra-rater reliability. For inter-rater analyses, only the first segmentation of the repeated brain was used.

Raters were asked to segment the following structures on one hemisphere of the brain: ACA; AHN; MV; RSP; SSp; SUB; VISp; VISam VPM and DG-sg. During the analysis, we found strong influence of the z-choice on the human DG-sg segmentations. We therefore excluded this structure from segmentation analysis (see Results). For each target structure, the task proceeded as follows: an STP stack consisting of 40 images (step size: 15 µm) was presented to the rater on a digitizer-pen-enabled monitor (Wacom Cintiq 22HD). The rater was also presented with a single plate from the online version of the Allen Mouse Brain Atlas on a second monitor and asked to outline the target structure in the STP data set. The raters were free to browse all Allen atlas plates to orient themselves along the anterior–posterior axis if necessary. The image stacks were presented using a custom Fiji/ImageJ[35,36] plugin that handled loading of images and logging of results. The order in which the brain structures were presented was random but identical for each participant within a group. Each set ($n = 4$) of 10 different structures was sampled from multiple brains.

Scoring and analysis. Manual segmentations were first manually cleaned by removing, for example, isolated touches that may appear when a rater accidentally clicks on an unrelated part of the data set in draw mode. From a total of 880 segmentations, we found five cases where the rater segmented the wrong structure or hemisphere. These cases were not included in the analyses. The remaining outlines of the segmented target structures were converted to filled binary images and downscaled to a pixel size of 4 µm in x–y. Due to the lack of a ground truth, all segmentations for a given target structure were compared to an 'consensus segmentation' derived from all manual segmentations of that structure using STAPLE. STAPLE is an iterative algorithm, designed to simultaneously assess the 'quality' of each segmentation and the average of all segmentations weighted by their quality. Quality is derived from the overlap of each segmentation with the agreement structure and is initialized to equal levels for all segmentations[20].

As an additional measure, we also calculated the inter-rater agreement using SBA[23], which gives the geometric mean of all segmentations. Both averaging methods yielded similar results (Supplementary Fig. 3a,b). Both the STAPLE and SBA consensus structure for each target were calculated using NiftySeg (seg_maths[37], https://sourceforge.net/projects/niftyseg). To score the quality of manual and automated segmentations, individual segmentations were compared to the consensus segmentation using the Dice score[21]. The Dice score is generally defined as the area of the intersection of two sets (that is, segmentations) divided by half the sum of the sets' areas and thus provides a measure of relative overlap between two segmentations. The Hausdorff metric was used as a supplementary scoring method (Supplementary Fig. 3b) and is generally defined as the longest distance between any point on one set and its closest neighbour in the other set and thus provides a good measure for the maximum distance between two segmentations.

Non-parametric tests were used to determine statistical significance, since data were found to be not normally distributed. Based on the observed effect sizes and number of repeats used for both manual and automated segmentations, such tests were performed with 100% power when the confidence interval was set to 99%.

Automated segmentation. The average brain data set from 3D mouse brain atlas by Kim *et al.* was aligned to the z-smoothed (Gaussian, 5 voxel s.d.) and then downsampled (12.5 µm per voxel isotropic) versions of the six brain data sets that had been used for manual segmentation. For registration, we used either the background fluorescence channel ($n = 5$) or a sparsely labelled RFP staining ($n = 1$). The first alignment step was an affine registration (NiftyReg, reg_aladin[34], six levels coarse-to-fine pyramidal approach of which the first five steps were computed) using a symmetric block-matching approach[34].

This was followed by a second free-form registration step, which places a regular grid of control points onto the reference image (NiftyReg, reg_f3d[19]). These control points are moved during registration, causing the surrounding image data to be moved, allowing for a local, non-linear alignment of the image data[38]. A parameter search was performed on two of the six brains to find suitable parameters for the free-form registration. Since image registration is a step-wise process that relies on assessing a cost function that embeds a measure of similarity between two data sets, we tested two similarity measures that both compare relative intensity differences in the atlas and the brain to be segmented: locally normalized cross-correlation and normalized mutual information[19]. Normalized mutual information, using 128 bins discretization achieved the highest overlap score and was hence used for aMAP. The remaining parameters achieving the highest overlap score were an initial Gaussian smoothing of the input images (with a 1 voxel s.d.), a control point grid spacing of 10 voxels isotropic, a bending energy weight of 0.95 and a six levels coarse-to-fine pyramidal approach of which only the first four steps were computed. For a more detailed description of the parameters, see the software manual distributed with aMAP. Unless specifically noted, the two brains on which the parameter search was performed were excluded from the analytical comparison of the manual versus automated segmentation.

The transformations obtained from registering the Kim *et al.* average brain to the individual image data sets were then applied to the 3D atlas from Kim *et al.* (Supplementary Fig. 1). Since this atlas is based on the original Allen brain atlas (generated from individual 2D segmentations of nissl-stained coronal plates[29]), structures show high-frequency fluctuations in border definition along the axis orthogonal to the atlas plane of section (Supplementary Fig. 6). To minimize their impact, the 3D atlas was smoothed twice using a Gaussian kernel with a s.d. of 0.5 voxels prior to being transformed (NiftySeg, seg_maths). Since the aMAP segmentations are 3D volumes, they cannot be directly compared with the 2D segmentations of human raters. Hence, the 3D volumes were converted to 2D outlines by making coronal sections through the part of the aMAP-generated 3D segmentation that corresponded to the stack given to human raters. The 2D outline with the highest Dice score was chosen as the result of the automated segmentation.

***z*-distance scoring.** To find the median distance between the sections chosen from the 40 optical sections in the STP data sets, each rater's z-choice was first determined and compared with the z-choice of all other raters segmenting the same STP data set. The absolute difference in section number for any two raters was then converted to distance by multiplying with the z-distance between two optical sections (15 µm). We calculated the absolute distance in z between two raters on the same brain and structure for all possible non-ordered combinations of raters.

Comparison of Dice- and landmark distance scoring. The following anatomical landmarks, as defined by the Waxholm space[27], (http://scalablebrainatlas.incf.org/main/coronal3d.php?template=WHS11&) and used in ref. 1, were placed in the average brain of the Kim *et al.* atlas and the downscaled version of each STP brain before registration: frontal middle 1; frontal right 2; frontal left 2; anterior commissure right; anterior commissure left; corpus callosum middle; hippocampus middle; interpeduncular nucleus right; interpeduncular nucleus middle; interpeduncular nucleus left. In one STP brain, hippocampus middle was omitted due to an imaging artefact in that region. The free-form registration was then rerun for each of the 6 brains using 18 different imposed bending energy (BE) weights (range: 0.2–0.99). The BE is used to penalize high-frequency transformation and acts as a regularization term in the optimization process. The optimization aims to find the best transformation parameters by maximizing the image similarity while minimizing the transformation BE. A BE weight that is set too low will lead to a mismatch of the segmentation due to artefacts caused by over-fitting the images. Setting the BE weight too high, on the other hand, will overly constrain the registration resulting in a more global mismatch between the segmentation outlines and the target brain. For the analysis of the suitability of the Dice score metric, the mean Dice score of all target structures in each brain (10 structures per brain) were plotted against the BE weight. Likewise, for the landmark distance analysis the mean distance between the landmarks in the brain data sets and the registered atlas were plotted against the BE weight.

Influence of *z*-range on Dice scores of human raters. To test whether the range of *z*-sections chosen by the human raters had a significant influence on the Dice scores of raters, we reanalysed a subset of the manual segmentations, choosing a window of three and seven consecutive *z*-sections for each structure in each brain. All segmentations that were not performed in this window were then discarded for this analysis. The position of the window was chosen on each brain and structure to contain the maximum possible number of human segmentations. New STAPLE consensus segmentations were generated for the *z*-limited analysis.

Brain structure schematic. To illustrate the position of the analysed structures and sections, 3D models were generated from the Allen Mouse Brain 3D voxel data using Fiji/ImageJ[35,36] to generate mesh models and blender (www.blender.org) to remesh, smooth and render them.

Data availability. Detailed instructions for setting up and using aMAP, including all necessary data and software, are openly available at http://www.swc.ucl.ac.uk/aMAP. This url also provides the published manual segmentations and validation pipeline and instructions on how to adapt the validation pipeline for other segmentation software.

References

1. Kim, Y. *et al.* Mapping social behavior-induced brain activation at cellular resolution in the mouse. *Cell Rep.* **10,** 292–305 (2015).
2. Oh, S. W. *et al.* A mesoscale connectome of the mouse brain. *Nature* **508,** 207–214 (2014).
3. Vélez-Fort, M. *et al.* The stimulus selectivity and connectivity of layer six principal cells reveals cortical microcircuits underlying visual processing. *Neuron* 1431–1443 (2014).
4. Li, N., Chen, T.-W., Guo, Z. V., Gerfen, C. R. & Svoboda, K. A motor cortex circuit for motor planning and movement. *Nature* **519,** 51–56 (2015).
5. Randlett, O. *et al.* Whole-brain activity mapping onto a zebrafish brain atlas. *Nat. Methods* **12,** 1039–1046 (2015).
6. Gong, H. *et al.* Continuously tracing brain-wide long-distance axonal projections in mice at a one-micron voxel resolution. *NeuroImage* **74,** 87–98 (2013).
7. Li, A. *et al.* Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain. *Science* **330,** 1404–1408 (2010).
8. Osten, P. & Margrie, T. W. Mapping brain circuitry with a light microscope. *Nat. Methods* **10,** 515–523 (2013).
9. Ragan, T. *et al.* Serial two-photon tomography for automated *ex vivo* mouse brain imaging. *Nat. Methods* **9,** 255–258 (2012).
10. Schwarz, M. K. *et al.* Fluorescent-protein stabilization and high-resolution imaging of cleared, intact mouse brains. *PloS ONE* **10,** e0124650 (2015).
11. Economo, M. N. *et al.* A platform for brain-wide imaging and reconstruction of individual neurons. *Elife* **5,** e10566 (2016).
12. Callaway, E. M. & Luo, L. Monosynaptic circuit tracing with glycoprotein-deleted rabies viruses. *J. Neurosci.* **35,** 8979–8985 (2015).
13. Murphey, D. K., Herman, A. M. & Arenkiel, B. R. Dissecting inhibitory brain circuits with genetically-targeted technologies. *Front. Neural Circuits* **8,** 124 (2014).
14. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445,** 168–176 (2007).
15. Niedworok, C. J. *et al.* Charting monosynaptic connectivity maps by two-color light-sheet fluorescence microscopy. *Cell Rep.* 1375–1386 (2012).
16. Vousden, D. A. *et al.* Whole-brain mapping of behaviourally induced neural activation in mice. *Brain. Struct. Funct.* **220,** 2043–2057 (2015).
17. Menegas, W. *et al.* Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. *Elife* **4,** e10032 (2015).
18. Klein, S., Staring, M., Murphy, K., Viergever, M. A. & Pluim, J. P. elastix: a toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* **29,** 196–205 (2010).
19. Modat, M. *et al.* Fast free-form deformation using graphics processing units. *Comput. Methods Programs Biomed.* **98,** 278–284 (2010).
20. Warfield, S. K., Zou, K. H. & Wells, W. M. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans. Med. Imaging* **23,** 903–921 (2004).
21. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26,** 297–302 (1945).
22. Leung, K. K. *et al.* Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. *NeuroImage* **51,** 1345–1359 (2010).
23. Rohlfing, T., Maurer, J. & Calvin, R. Shape-based averaging. *IEEE Trans. Image Process.* **16,** 153–161 (2007).
24. Ou, Y., Akbari, H., Bilello, M., Da, X. & Davatzikos, C. Comparative evaluation of registration algorithms in different brain databases with varying difficulty: results and insights. *IEEE Trans. Med. Imaging* **33,** 2039–2065 (2014).
25. Rohlfing, T. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Trans. Med. Imaging* **31,** 153–163 (2012).
26. Dorr, A. E., Lerch, J. P., Spring, S., Kabani, N. & Henkelman, R. M. High resolution three-dimensional brain atlas using an average magnetic resonance image of 40 adult C57Bl/6J mice. *NeuroImage* **42,** 60–69 (2008).
27. Johnson, G., Badea, A. & Brandenburg, J. Waxholm space: an image-based reference for coordinating mouse brain research. *NeuroImage* **53,** 365–372 (2010).
28. Ma, Y. *et al.* A three-dimensional digital atlas database of the adult C57BL/6J mouse brain by magnetic resonance microscopy. *Neuroscience* **135,** 1203–1215 (2005).
29. Ng, L. *et al.* Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **4,** 382–393 (2007).
30. Barnes, J. *et al.* A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. *NeuroImage* **40,** 1655–1671 (2008).
31. Artaechevarria, X., Munoz-Barrutia, A. & Ortiz-de-Solorzano, C. Combination strategies in multi-atlas image segmentation: application to brain MR data. *IEEE Trans. Med. Imaging* **28,** 1266–1277 (2009).
32. Iglesias, J. E. & Sabuncu, M. R. Multi-atlas segmentation of biomedical images: a survey. *Med. Image Anal.* **24,** 205–219 (2015).
33. Ma, D. *et al.* Automatic structural parcellation of mouse brain MRI using multi-atlas label fusion. *PloS ONE* **9,** e86576 (2014).
34. Modat, M. *et al.* Global image registration using a symmetric block-matching approach. *J. Med. Imaging (Bellingham)* **1,** 024003 (2014).
35. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9,** 676–682 (2012).
36. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9,** 671–675 (2012).
37. Jorge Cardoso, M. *et al.* STEPS: similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. *Med. Image Anal.* **17,** 671–684 (2013).
38. Rueckert, D. *et al.* Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans. Med. Imaging* **18,** 712–721 (1999).

Acknowledgements

The authors thank members of the former Division of Neurophysiology at the National Institute for Medical Research, who participated as human raters on the manual segmentation task. We also thank James Nelson and James Briscoe for input throughout the project. S.O. is supported by the EPSRC (EP/H046410/1, EP/J020990/1, EP/K005278), the MRC (MR/J01107X/1), the EU-FP7 project VPH-DARE@IT (FP7-ICT-2011-9-601055), the NIHR Biomedical Research Unit (Dementia) at UCL and the National Institute for Health Research University College London Hospitals Biomedical Research Centre (NIHR BRC UCLH/UCL High Impact Initiative- BW.mn. BRC10269). M.M. is supported by the UCL Leonard Wolfson Experimental Neurology Centre (PR/ylr/18575). This project was funded by the Medical Research Council (MC_U1175975156), The Gatsby Charitable Trust and The Wellcome Trust (WT096436AIA) (T.W.M.).

Author contributions

C.J.N. and T.W.M. conceived the project. C.J.N., J.M.C., S.O. and M.M. performed software development for aMAP. C.J.N. implemented aMAP and analysed segmentation

performance with input from all other authors. P.O. provided the atlas. C.J.N. and T.W.M. wrote the paper with input from all other authors.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no conflict of interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Niedworok, C. J. *et al.* aMAP is a validated pipeline for registration and segmentation of high-resolution mouse brain data. *Nat. Commun.* 7:11879 doi: 10.1038/ncomms11879 (2016).

Reference List

1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Abbas, S.S., Dijkstra, T.M., and Heskes, T. (2014). A comparative study of cell classifiers for image-based high-throughput screening. BMC Bioinformatics *15*, 342.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding *110*, 346-359.

Belforte, J.E., Zsiros, V., Sklar, E.R., Jiang, Z., Yu, G., Li, Y., Quinlan, E.M., and Nakazawa, K. (2010). Postnatal NMDA receptor ablation in corticolimbic interneurons confers schizophrenia-like phenotypes. Nature neuroscience *13*, 76-83.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. Advances in neural information processing systems *19*, 153.

Bennett, G., Giamberardino, L.D., Koenig, H., and Droz, B. (1973). Axonal migration of protein and glycoprotein to nerve endings. II. Radioautographic analysis of the renewal of glycoproteins in nerve endings of chicken ciliary ganglion after intracerebral injection of [3H]fucose and [3H]glucosamine. Brain Research *60*, 129-146.

Bishop, C.M. (2006). Pattern recognition and machine learning (New York: Springer).

Biswal, B., Yetkin, F.Z., Haughton, V.M., and Hyde, J.S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med *34*, 537-541.

Bliss, T.V., and Collingridge, G.L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. Nature *361*, 31-39.

Bohland, J.W., Wu, C., Barbas, H., Bokil, H., Bota, M., Breiter, H.C., Cline, H.T., Doyle, J.C., Freed, P.J., Greenspan, R.J.*, et al.* (2009). A proposal for a coordinated effort for the determination of brainwide neuroanatomical connectivity in model organisms at a mesoscopic scale. PLoS computational biology *5*, e1000334.

Bolte, S., and Cordelieres, F.P. (2006). A guided tour into subcellular colocalization analysis in light microscopy. Journal of Microscopy-Oxford *224*, 213-232.

Bottou, L. (2012). Stochastic gradient descent tricks. In Neural Networks: Tricks of the Trade (Springer), pp. 421-436.

Breiman, L. (1996). Bagging predictors. Machine learning *24*, 123-140.

Broyd, S.J., Demanuele, C., Debener, S., Helps, S.K., James, C.J., and Sonuga-Barke, E.J. (2009). Default-mode brain dysfunction in mental disorders: a systematic review. Neuroscience & biobehavioral reviews *33*, 279-296.

Büttner-Ennever, J.A., Grob, P., Akert, K., and Bizzini, B. (1981). A Transsynaptic Autoradiographic Study of the Pathways Controlling the Extraocular Eye Muscles, using [125I]B-IIb Tetanus Toxin Fragment. Annals of the New York Academy of Sciences *374*, 157-170.

Cajal, S.R.y. (1894). Les nouvelles idées sur la structure du système nerveux chez l'homme et chez les vertébrés. 200.

Cajal, S.R.y. (1896). Beitrag zum Studium der Medulla Oblongata, des Kleinhirns und des Ursprungs der Gehirnnerven. 139.

Cajal, S.R.y. (1899). Comparative study of the sensory areas of the human cortex. 72.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J.*, et al.* (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol *7*, R100.

Chklovskii, D.B. (2004). Synaptic connectivity and neuronal morphology: two sides of the same coin. Neuron *43*, 609-617.

Cho, Y., and Saul, L.K. (2009). Kernel methods for deep learning. Paper presented at: Advances in neural information processing systems.

Chung, K., Wallace, J., Kim, S.-Y., Kalyanasundaram, S., Andalman, A.S., Davidson, T.J., Mirzabekov, J.J., Zalocusky, K.a., Mattis, J., Denisin, A.K.*, et al.* (2013). Structural and molecular interrogation of intact biological systems. Nature *497*, 332-337.

Collins, D.L., Holmes, C.J., Peters, T.M., and Evans, A.C. (1995). Automatic 3-D model-based neuroanatomical segmentation. Human Brain Mapping *3*, 190-208.

Contestabile, A., Benfenati, F., and Gasparini, L. (2010). Communication breaks-Down: from neurodevelopment defects to cognitive disabilities in Down syndrome. Progress in neurobiology *91*, 1-22.

Cowan, W.M. (1998). The emergence of modern neuroanatomy and developmental neurobiology. Neuron *20*, 413-426.

Denk, W., and Horstmann, H. (2004). Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. PLoS biology *2*, e329.

Desai, M., Kahn, I., Knoblich, U., Bernstein, J., Atallah, H., Yang, A., Kopell, N., Buckner, R.L., Graybiel, A.M., and Moore, C.I. (2011). Mapping brain networks in awake mice using combined optical neural control and fMRI. Journal of neurophysiology *105*, 1393-1405.

Dice, L.R. (1945). Measures of the Amount of Ecologic Association Between Species. Ecology *26*, 297-302.

Dodt, H.U.H.-U., Leischner, U., Schierloh, A., Jährling, N., Mauch, C.P.C.P., Deininger, K., Deussing, J.M.J.M., Eder, M., Zieglgänsberger, W., and Becker, K. (2007). Ultramicroscopy: three-dimensional visualization of neuronal networks in the whole mouse brain. Nature methods *4*, 331-336.

Economo, M.N., Clack, N.G., Lavis, L.D., Gerfen, C.R., Svoboda, K., Myers, E.W., and Chandrashekar, J. (2016). A platform for brain-wide imaging and reconstruction of individual neurons. Elife *5*.

Etessami, R., Conzelmann, K.K., Fadai-Ghotbi, B., Natelson, B., Tsiang, H., and Ceccaldi, P.E. (2000). Spread and pathogenic characteristics of a G-deficient rabies virus recombinant: an in vitro and in vivo study. J Gen Virol *81*, 2147-2153.

Farley, B., and Clark, W. (1954). Simulation of self-organizing systems by digital computer. Transactions of the IRE Professional Group on Information Theory *4*, 76-84.

Feigenbaum, E.A., and Feldman, J. (1963). Computers and thought, a collection of articles by Armer [and others] (New York,: McGraw-Hill).

Fu, Y., Tucciarone, J.M., Espinosa, J.S., Sheng, N., Darcy, D.P., Nicoll, R.A., Huang, Z.J., and Stryker, M.P. (2014). A cortical circuit for gain control by behavioral state. Cell *156*, 1139-1152.

Geschwind, D.H., and Levitt, P. (2007). Autism spectrum disorders: developmental disconnection syndromes. Current opinion in neurobiology *17*, 103-111.

Gillet, J.P., Derer, P., and Tsiang, H. (1986). Axonal transport of rabies virus in the central nervous system of the rat. Journal of Neuropathology & Experimental Neurology *45*, 619-634.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., *et al.* (2013). The minimal preprocessing pipelines for the Human Connectome Project. NeuroImage *80*, 105-124.

Golgi, C. (1875). Sulla fina struttura dei bulbi olfattorii.

Golgi, C. (1886). Sulla fina anatomia degli organi centrali del sistema nervoso.

Golgi, C. (1989). On the structure of nerve cells. 1898. Journal of microscopy *155*, 3-7.

Gong, H., Zeng, S., Yan, C., Lv, X., Yang, Z., Xu, T., Feng, Z., Ding, W., Qi, X., Li, A., *et al.* (2013). Continuously tracing brain-wide long-distance axonal projections in mice at a one-micron voxel resolution. NeuroImage *74*, 87-98.

Gordon, J.W., and Ruddle, F.H. (1981). Integration and stable germ line transmission of genes injected into mouse pronuclei. Science *214*, 1244-1246.

Götz, J., and Ittner, L.M. (2008). Animal models of Alzheimer's disease and frontotemporal dementia. Nature Reviews Neuroscience *9*, 532-544.

Haller, J.W., Banerjee, A., Christensen, G.E., Gado, M., Joshi, S., Miller, M.I., Sheline, Y., Vannier, M.W., and Csernansky, J.G. (1997). Three-dimensional hippocampal MR morphometry with high-dimensional transformation of a neuroanatomic atlas. Radiology *202*, 504-510.

Hama, H., Kurokawa, H., Kawano, H., Ando, R., Shimogori, T., Noda, H., Fukami, K., Sakaue-Sawano, A., and Miyawaki, A. (2011). Scale: a chemical approach for fluorescence imaging and reconstruction of transparent mouse brain. Nature neuroscience *14*, 1481-1488.

Han, J.W., Breckon, T.P., Randell, D.A., and Landini, G. (2012). The application of support vector machine classification to detect cell nuclei for automated microscopy. Machine Vision and Applications *23*, 15-24.

Harris, J.A., Hirokawa, K.E., Sorensen, S.A., Gu, H., Mills, M., Ng, L.L., Bohn, P., Mortrud, M., Ouellette, B., and Kidney, J. (2015). Anatomical characterization of Cre driver mice for neural circuit mapping and manipulation. Neural Circuits Revealed, 71.

He, K., Zhang, X., Ren, S., and Sun, J. (2015a). Deep Residual Learning for Image Recognition. ArxivOrg *7*, 171-180.

He, Y., Gong, H., Xiong, B., Xu, X., Li, A., Jiang, T., Sun, Q., Wang, S., Luo, Q., and Chen, S. (2015b). iCut: an Integrative Cut Algorithm Enables Accurate Segmentation of Touching Cells. Sci Rep *5*, 12089.

Hearst, M.A., Dumais, S.T., Osman, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their Applications *13*, 18-28.

Helmstaedter, M. (2013). Cellular-resolution connectomics: challenges of dense neural circuit reconstruction. Nature methods *10*, 501-507.

Helmstaedter, M., Briggman, K.L., and Denk, W. (2008). 3D structural imaging of the brain with photons and electrons. Current opinion in neurobiology *18*, 633-641.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. ArXiv e-prints, 1-18.

Hunnicutt, B.J., Long, B.R., Kusefoglu, D., Gertz, K.J., Zhong, H., and Mao, T. (2014). A comprehensive thalamocortical projection map at the mesoscopic level. Nature neuroscience, 1-13.

Iandola, F.N., Ashraf, K., Moskewicz, M.W., and Keutzer, K. (2015). FireCaffe: near-linear acceleration of deep neural network training on compute clusters. arXiv preprint arXiv:151100175.

Irshad, H., Veillard, A., Roux, L., and Racoceanu, D. (2014). Methods for Nuclei Detection, Segmentation, and Classification in Digital Histopathology: A Review - Current Status and Future Potential. IEEE Reviews in Biomedical Engineering *7*, 97-114.

Jankowska, E. (1985). Further indications for enhancement of retrograde transneuronal transport of WGA-HRP by synaptic activity. Brain research *341*, 403-408.

Jansen, A.S., Nguyen, X.V., Karpitskiy, V., Mettenleiter, T.C., and Loewy, A.D. (1995). Central command neurons of the sympathetic nervous system: basis of the fight-or-flight response. Science (New York, NY) *270*, 644-646.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. MM '14 Proceedings of the 22nd ACM international conference on Multimedia, 675-678.

Johnson, G., Badea, A., and Brandenburg, J. (2010). Waxholm Space: An image-based reference for coordinating mouse brain research. NeuroImage *53*, 365-372.

Jones, E.G., and Hartman, B.K. (1978). Recent advances in neuroanatomical methodology. Annual review of neuroscience *1*, 215-296.

Jorge Cardoso, M., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N.C., Ourselin, S., and Alzheimer's Disease Neuroimaging, I. (2013). STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation. Medical image analysis *17*, 671-684.

Josh Huang, Z., and Zeng, H. (2013). Genetic approaches to neural circuits in the mouse. Annual review of neuroscience *36*, 183-215.

Kalueff, A., Wheaton, M., and Murphy, D. (2007). What's wrong with my mouse model?: Advances and strategies in animal modeling of anxiety and depression. Behavioural brain research *179*, 1-18.

Kasthuri, N., Hayworth, K.J., Berger, D.R., Schalek, R.L., Conchello, J.A., Knowles-Barley, S., Lee, D., Vázquez-Reina, A., Kaynig, V., and Jones, T.R. (2015). Saturated reconstruction of a volume of neocortex. Cell *162*, 648-661.

Katz, L.C., Burkhalter, A., and Dreyer, W.J. (1984). Fluorescent latex microspheres as a retrograde neuronal marker for in vivo and in vitro studies of visual cortex. Nature *310*, 498-500.

Kim, E.J., Jacobs, M.W., Ito-Cole, T., and Callaway, E.M. (2016). Improved Monosynaptic Neural Circuit Tracing Using Engineered Rabies Virus Glycoproteins. Cell reports.

Kim, J.S., Greene, M.J., Zlateski, A., Lee, K., Richardson, M., Turaga, S.C., Purcaro, M., Balkam, M., Robinson, A., Behabadi, B.F.*, et al.* (2014a). Space-time wiring specificity supports direction selectivity in the retina. Nature *509*, 331-336.

Kim, Y., Venkataraju, K.U., Pradhan, K., Mende, C., Taranda, J., Turaga, S.C., Arganda-Carreras, I., Ng, L., Hawrylycz, M.J., Rockland, K.S.*, et al.* (2014b). Mapping Social Behavior-Induced Brain Activation at Cellular Resolution in the Mouse. Cell reports, 1-14.

Klein, F., Mouquet, H., Dosenovic, P., Scheid, J.F., Scharf, L., and Nussenzweig, M.C. (2013). Antibodies in HIV-1 vaccine development and therapy. Science *341*, 1199-1204.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., and Pluim, J.P. (2010). elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging *29*, 196-205.

Kristensson, K., Nennesmo, I., Persson, L., and Lycke, E. (1982). Neuron to neuron transmission of herpes simplex virus: Transport of virus from skin to brainstem nuclei. Journal of the Neurological Sciences *54*, 149-156.

Kristensson, K., Olsson, Y., and Sjöstrand, J. (1971). Axonal uptake and retrograde transport of exogenous proteins in the hypoglossal nerve. Brain research *32*, 399-406.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. Paper presented at: Advances in neural information processing systems.

Kuypers, H.G., and Ugolini, G. (1990). Viruses as transneuronal tracers. Trends in neurosciences *13*, 71-75.

Lammel, S., Lim, B.K., Ran, C., Huang, K.W., Betley, M.J., Tye, K.M., Deisseroth, K., and Malenka, R.C. (2012). Input-specific control of reward and aversion in the ventral tegmental area. Nature *491*, 212-217.

LaTorre, A., Alonso-Nanclares, L., Muelas, S., Peña, J.-M., and DeFelipe, J. (2013). 3D segmentations of neuronal nuclei from confocal microscope image stacks. Frontiers in neuroanatomy *7*.

Le Cun, B.B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., and Jackel, L.D. (1990). Handwritten digit recognition with a back-propagation network. Paper presented at: Advances in neural information processing systems (Citeseer).

Lee, E., Choi, J., Jo, Y., Kim, J.Y., Jang, Y.J., Lee, H.M., Kim, S.Y., Lee, H.J., Cho, K., Jung, N.*, et al.* (2016). ACT-PRESTO: Rapid and consistent tissue clearing and labeling method for 3-dimensional (3D) imaging. Sci Rep *6*, 18631.

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J.*, et al.* (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature *445*, 168-176.

Leung, K.K., Barnes, J., Ridgway, G.R., Bartlett, J.W., Clarkson, M.J., Macdonald, K., Schuff, N., Fox, N.C., Ourselin, S., and Alzheimer's Disease Neuroimaging, I. (2010).

Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. NeuroImage *51*, 1345-1359.

Lewis, B.C., Chinnasamy, N., Morgan, R.A., and Varmus, H.E. (2001). Development of an avian leukosis-sarcoma virus subgroup A pseudotyped lentiviral vector. Journal of virology *75*, 9339-9344.

Li, Z., Yu, T., Morishima, M., Pao, A., LaDuca, J., Conroy, J., Nowak, N., Matsui, S., Shiraishi, I., and Yu, Y.E. (2007). Duplication of the entire 22.9 Mb human chromosome 21 syntenic region on mouse chromosome 16 causes cardiovascular and gastrointestinal abnormalities. Hum Mol Genet *16*, 1359-1366.

Lo, L., and Anderson, D.J. (2011). A cre-dependent, anterograde transsynaptic viral tracer for mapping output pathways of genetically marked neurons. Neuron *72*, 938-950.

Lowe, D.G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision *60*, 91-110.

Luck, S.J., Chelazzi, L., Hillyard, S.A., and Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. Journal of neurophysiology *77*, 24-42.

Lundh, B. (1990). Spread of vesicular stomatitis virus along the visual pathways after retinal infection in the mouse. Acta Neuropathol *79*, 395-401.

Lundh, B., Kristensson, K., and Norrby, E. (1987). Selective Infections of Olfactory and Respiratory Epithelium by Vesicular Stomatitis and Sendai Viruses. Neuropathology and Applied Neurobiology, 111-122.

Luo, L., Callaway, E.M., and Svoboda, K. (2008). Genetic dissection of neural circuits. Neuron *57*, 634-660.

Ma, D., Cardoso, M.J., Modat, M., Powell, N., Wells, J., Holmes, H., Wiseman, F., Tybulewicz, V., Fisher, E., Lythgoe, M.F*., et al.* (2014). Automatic structural parcellation of mouse brain MRI using multi-atlas label fusion. PloS one *9*, e86576.

Ma, Y., Hof, P.R., Grant, S.C., Blackband, S.J., Bennett, R., Slatest, L., McGuigan, M.D., and Benveniste, H. (2005). A three-dimensional digital atlas database of the adult C57BL/6J mouse brain by magnetic resonance microscopy. Neuroscience *135*, 1203-1215.

Ma, Y., Smith, D., Hof, P.R., Foerster, B., Hamilton, S., Blackband, S.J., Yu, M., and Benveniste, H. (2008). In Vivo 3D Digital Atlas Database of the Adult C57BL/6J Mouse Brain by Magnetic Resonance Microscopy. Frontiers in neuroanatomy *2*, 1.

Madabhushi, A., and Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical image analysis *33*, 170-175.

Mallard, J.R. (2003). The evolution of medical imaging: from Geiger counters to MRI--a personal saga. Perspect Biol Med *46*, 349-370.

Malpica, N., de Solorzano, C.O., Vaquero, J.J., Santos, A., Vallcorba, I., Garcia-Sagredo, J.M., and del Pozo, F. (1997). Applying watershed algorithms to the segmentation of clustered nuclei. Cytometry *28*.

Marblestone, A.H., Zamft, B.M., Maguire, Y.G., Shapiro, M.G., Cybulski, T.R., Glaser, J.I., Amodei, D., Stranges, P.B., Kalhor, R., Dalrymple, D.a.*, et al.* (2013). Physical principles for scalable neural recording. Frontiers in computational neuroscience *7*, 137.

Marder, E., and Goaillard, J.M. (2006). Variability, compensation and homeostasis in neuron and network function. Nature reviews Neuroscience *7*, 563-574.

Marshel, J.H., Mori, T., Nielsen, K.J., and Callaway, E.M. (2010). Targeting single neuronal networks for gene expression and cell labeling in vivo. Neuron *67*, 562-574.

Martin, X., and Dolivo, M. (1983). Neuronal and transneuronal tracing in the trigeminal system of the rat using the herpes virus suis. Brain research *273*, 253-276.

Mebatsion, T., König, M., and Conzelmann, K.-K. (1996). Budding of Rabies Virus Particles in the Absence of the Spike Glycoprotein. Cell *84*, 941-951.

Meijering, E. (2012). Cell Segmentation: 50 Years Down the Road [Life Sciences]. IEEE Signal Processing Magazine *29*, 140-145.

Menegas, W., Bergan, J.F., Ogawa, S.K., Isogai, Y., Umadevi Venkataraju, K., Osten, P., Uchida, N., and Watabe-Uchida, M. (2015). Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. Elife *4*.

Meyer, F., and Beucher, S. (1990). Morphological segmentation. J Visual Communication Image Representation *1*.

Minsky, M., and Papert, S. (1969). Perceptrons; an introduction to computational geometry (Cambridge, Mass.,: MIT Press).

Misselwitz, B., Strittmatter, G., Periaswamy, B., Schlumberger, M.C., Rout, S., Horvath, P., Kozak, K., and Hardt, W.-D. (2010). Enhanced CellClassifier: a multi-class classification tool for microscopy images. BMC bioinformatics *11*, 30.

Miyamichi, K., Amat, F., Moussavi, F., Wang, C., Wickersham, I., Wall, N.R., Taniguchi, H., Tasic, B., Huang, Z.J., He, Z.*, et al.* (2011). Cortical representations of olfactory input by trans-synaptic tracing. Nature *472*, 191-196.

Modat, M., Cash, D.M., Daga, P., Winston, G.P., Duncan, J.S., and Ourselin, S. (2014). Global image registration using a symmetric block-matching approach. J Med Imaging (Bellingham) *1*, 024003.

Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., and Ourselin, S. (2010). Fast free-form deformation using graphics processing units. Computer methods and programs in biomedicine *98*, 278-284.

Morgan, J.L., and Lichtman, J.W. (2013). Why not connectomics? Nature methods *10*, 494-500.

Nair, V., and Hinton, G.E. (2010). Rectified linear units improve restricted boltzmann machines. Paper presented at: Proceedings of the 27th International Conference on Machine Learning (ICML-10).

Nestler, E.J., and Hyman, S.E. (2010). Animal models of neuropsychiatric disorders. Nature neuroscience *13*, 1161-1169.

Nestor, S.M., Gibson, E., Gao, F.Q., Kiss, A., Black, S.E., and Alzheimer's Disease Neuroimaging, I. (2013). A direct morphometric comparison of five labeling protocols for multi-atlas driven automatic segmentation of the hippocampus in Alzheimer's disease. NeuroImage *66*, 50-70.

Ng, L., Pathak, S.D., Kuan, C., Lau, C., Dong, H., Sodt, A., Dang, C., Avants, B., Yushkevich, P., Gee, J.C.*, et al.* (2007). Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM *4*, 382-393.

Nickolls, J., Buck, I., Garland, M., and Skadron, K. (2008). Scalable parallel programming with CUDA. Queue *6*, 40-53.

Niedworok, C.J., Brown, A.P., Jorge Cardoso, M., Osten, P., Ourselin, S., Modat, M., and Margrie, T.W. (2016). aMAP is a validated pipeline for registration and segmentation of high-resolution mouse brain data. Nat Commun *7*, 11879.

Niedworok, C.J., Schwarz, I., Ledderose, J., Giese, G., Conzelmann, K.-K., and Schwarz, M.K. (2012). Charting Monosynaptic Connectivity Maps by Two-Color Light-Sheet Fluorescence Microscopy. Cell reports, 1375-1386.

Oberlaender, M., Dercksen, V.J., Egger, R., Gensel, M., Sakmann, B., and Hege, H.-C. (2009). Automated three-dimensional detection and counting of neuron somata. Journal of neuroscience methods *180*, 147-160.

Oh, S.W., Harris, J.a., Ng, L., Winslow, B., Cain, N., Mihalas, S., Wang, Q., Lau, C., Kuan, L., Henry, A.M.*, et al.* (2014). A mesoscale connectome of the mouse brain. Nature *508*, 207-214.

Orban, P.C., Chui, D., and Marth, J.D. (1992). Tissue-and site-specific DNA recombination in transgenic mice. Proceedings of the National Academy of Sciences *89*, 6861-6865.

Osakada, F., Mori, T., Cetin, A.H., Marshel, J.H., Virgen, B., and Callaway, E.M. (2011). New rabies virus variants for monitoring and manipulating activity and gene expression in defined neural circuits. Neuron *71*, 617-631.

Osten, P., and Margrie, T.W. (2013). Mapping brain circuitry with a light microscope. Nature methods *10*, 515-523.

Otsu, N. (1975). A threshold selection method from gray-level histograms. Automatica *11*, 23-27.

Ou, Y., Akbari, H., Bilello, M., Da, X., and Davatzikos, C. (2014). Comparative evaluation of registration algorithms in different brain databases with varying difficulty: results and insights. IEEE transactions on medical imaging *33*, 2039-2065.

Paxinos, G., and Franklin, K.B.J. (2004). The Mouse Brain in Stereotaxic Coordinates.

Peça, J., Feliciano, C., Ting, J.T., Wang, W., Wells, M.F., Venkatraman, T.N., Lascola, C.D., Fu, Z., and Feng, G. (2011). Shank3 mutant mice display autistic-like behaviours and striatal dysfunction. Nature *472*, 437-442.

Peng, H., Bria, A., Zhou, Z., Iannello, G., and Long, F. (2014). Extensible visualization and analysis for multidimensional images using Vaa3D. Nature protocols *9*, 193-208.

Peng, H., Hawrylycz, M., Roskams, J., Hill, S., Spruston, N., Meijering, E., and Ascoli, G.A. (2015). BigNeuron: Large-Scale 3D Neuron Reconstruction from Optical Microscopy Images. Neuron *87*, 252-256.

Pietzsch, T., Saalfeld, S., Preibisch, S., and Tomancak, P. (2015). BigDataViewer: visualization and processing for large image data sets. Nature methods *12*, 481-483.

Preibisch, S., Saalfeld, S., and Tomancak, P. (2009). Globally optimal stitching of tiled 3D microscopic image acquisitions. Bioinformatics *25*, 1463-1465.

Przedborski, S., and Vila, M. (2003). The 1‐Methyl‐4‐Phenyl‐1, 2, 3, 6‐Tetrahydropyridine Mouse Model. Annals of the New York Academy of Sciences *991*, 189-198.

Quinlan, J.R. (1986). Induction of Decision Trees. Machine Learning *1*, 81-106.

Ragan, T., Kadiri, L.R., Venkataraju, K.U., Bahlmann, K., Sutin, J., Taranda, J., Arganda-Carreras, I., Kim, Y., Seung, H.S., and Osten, P. (2012). Serial two-photon tomography for automated ex vivo mouse brain imaging. Nature methods.

Rancz, E.A., Franks, K.M., Schwarz, M.K., Pichler, B., Schaefer, A.T., and Margrie, T.W. (2011). Transfection via whole-cell recording in vivo: bridging single-cell physiology, genetics and connectomics. Nature neuroscience *14*, 527-532.

Reardon, T.R., Murray, A.J., Turi, G.F., Wirblich, C., Croce, K.R., Schnell, M.J., Jessell, T.M., and Losonczy, A. (2016). Rabies Virus CVS-N2c(DeltaG) Strain Enhances Retrograde Synaptic Transfer and Neuronal Viability. Neuron *89*, 711-724.

Reid, R.C. (2012). From functional architecture to functional connectomics. Neuron *75*, 209-217.

Renier, N., Adams, E.L., Kirst, C., Wu, Z., Azevedo, R., Kohl, J., Autry, A.E., Kadiri, L., Umadevi Venkataraju, K., Zhou, Y.*, et al.* (2016). Mapping of Brain Activity by Automated Volume Analysis of Immediate Early Genes. Cell *165*, 1789-1802.

Richards, K., Watson, C., Buckley, R.F., Kurniawan, N.D., Yang, Z., Keller, M.D., Beare, R., Bartlett, P.F., Egan, G.F., Galloway, G.J.*, et al.* (2011). Segmentation of the mouse hippocampal formation in magnetic resonance images. NeuroImage *58*, 732-740.

Richardson, D.S., and Lichtman, J.W. (2015). Clarifying Tissue Clearing. Cell *162*, 246-257.

Rochester, N., Holland, J., Haibt, L., and Duda, W. (1956). Tests on a cell assembly theory of the action of the brain, using a large digital computer. IRE Transactions on Information Theory *2*, 80-93.

Rohlfing, T. (2012). Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. IEEE transactions on medical imaging *31*, 153-163.

Rohlfing, T., and Maurer, J., Calvin R. (2007). Shape-Based Averaging. IEEE Transactions on Image Processing *16*, 153-161.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review *65*, 386.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.*, et al.* (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision *115*, 211-252.

Sarle, W.S. (1996). Stopped training and other remedies for overfitting. Computing science and statistics, 352-360.

Sauer, B., and Henderson, N. (1988). Site-specific DNA recombination in mammalian cells by the Cre recombinase of bacteriophage P1. Proceedings of the National Academy of Sciences *85*, 5166-5170.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks *61*, 85-117.

Schneider, C.a., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nature methods *9*, 671-675.

Schwarz, M.K., Scherbarth, A., Sprengel, R., Engelhardt, J., Theer, P., and Giese, G. (2015). Fluorescent-protein stabilization and high-resolution imaging of cleared, intact mouse brains. PloS one *10*, e0124650.

Semenza, G.L. (2003). Targeting HIF-1 for cancer therapy. Nature reviews cancer *3*, 721-732.

Sigurdsson, T., Stark, K.L., Karayiorgou, M., Gogos, J.A., and Gordon, J.A. (2010). Impaired hippocampal-prefrontal synchrony in a genetic mouse model of schizophrenia. Nature *464*, 763-767.

Silvestri, L., Paciscopi, M., Soda, P., Biamonte, F., Iannello, G., Frasconi, P., and Pavone, F.S. (2015). Quantitative neuroanatomy of all Purkinje cells with light sheet microscopy and high-throughput image analysis. Frontiers in neuroanatomy *9*, 1-11.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556.

Sjostrand, F.S. (1958). Ultrastructure of retinal rod synapses of the guinea pig eye as revealed by three-dimensional reconstructions from serial sections. J Ultrastruct Res *2*, 122-170.

Sommer, C., Straehle, C., Köthe, U., and Hamprecht, F.A. (2011). Ilastik: Interactive learning and segmentation toolkit. Paper presented at: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro.

Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: A structural description of the human brain. PLoS computational biology *1*, e42.

Stone, J.E., Gohara, D., and Shi, G. (2010). OpenCL: A parallel programming standard for heterogeneous computing systems. Computing in science & engineering *12*, 66-73.

Sun, N., Cassell, M.D., and Perlman, S. (1996). Anterograde, transneuronal transport of herpes simplex virus type 1 strain H129 in the murine visual system. Journal of virology *70*, 5405-5413.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Szigeti, B., Gleeson, P., Vella, M., Khayrulin, S., Palyanov, A., Hokanson, J., Currie, M., Cantarelli, M., Idili, G., and Larson, S. (2014). OpenWorm: an open-science approach to modeling Caenorhabditis elegans. Frontiers in computational neuroscience *8*, 137.

Toyoshima, Y., Tokunaga, T., Hirose, O., Kanamori, M., Teramoto, T., Jang, M.S., Kuge, S., Ishihara, T., Yoshida, R., and Iino, Y. (2016). Accurate Automatic Detection of Densely Distributed Cell Nuclei in 3D Space. PLoS computational biology *12*, e1004970.

Ugolini, G. (1995). Specificity of rabies virus as a transneuronal tracer of motor networks: transfer from hypoglossal motoneurons to connected second-order and higher order central nervous system cell groups. J Comp Neurol *356*, 457-480.

Uhr, L., and Vossler, C. (1961). A pattern recognition program that generates, evaluates, and adjusts its own operators. Paper presented at: Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference (ACM).

Ullmann, J.F., Keller, M.D., Watson, C., Janke, A.L., Kurniawan, N.D., Yang, Z., Richards, K., Paxinos, G., Egan, G.F., Petrou, S.*, et al.* (2012). Segmentation of the

C57BL/6J mouse cerebellum in magnetic resonance images. NeuroImage *62*, 1408-1414.

Ullmann, J.F., Watson, C., Janke, A.L., Kurniawan, N.D., Paxinos, G., and Reutens, D.C. (2014). An MRI atlas of the mouse basal ganglia. Brain structure & function *219*, 1343-1353.

Ullmann, J.F., Watson, C., Janke, A.L., Kurniawan, N.D., and Reutens, D.C. (2013). A segmentation protocol and MRI atlas of the C57BL/6J mouse neocortex. NeuroImage *78*, 196-203.

Vélez-Fort, M., Rousseau, Charly V., Niedworok, Christian J., Wickersham, Ian R., Rancz, Ede A., Brown, Alexander P.Y., Strom, M., and Margrie, Troy W. (2014). The Stimulus Selectivity and Connectivity of Layer Six Principal Cells Reveals Cortical Microcircuits Underlying Visual Processing. Neuron, 1431-1443.

Vousden, D.A., Epp, J., Okuno, H., Nieman, B.J., van Eede, M., Dazai, J., Ragan, T., Bito, H., Frankland, P.W., Lerch, J.P*., et al.* (2015). Whole-brain mapping of behaviourally induced neural activation in mice. Brain structure & function *220*, 2043-2057.

Wall, N.R., Wickersham, I.R., Cetin, A., De La Parra, M., and Callaway, E.M. (2010). Monosynaptic circuit tracing in vivo through Cre-dependent targeting and complementation of modified rabies virus. Proceedings of the National Academy of Sciences of the United States of America *107*, 21848-21853.

Warfield, S.K., Zou, K.H., and Wells, W.M. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging *23*, 903-921.

Wenger, N., Moraud, E.M., Raspopovic, S., Bonizzato, M., DiGiovanna, J., Musienko, P., Morari, M., Micera, S., and Courtine, G. (2014). Closed-loop neuromodulation of spinal sensorimotor circuits controls refined locomotion after complete spinal cord injury. Science translational medicine *6*, 255ra133-255ra133.

Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences.

White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. Philos Trans R Soc Lond B Biol Sci *314*, 1-340.

Wickersham, I.R., Finke, S., Conzelmann, K.K., and Callaway, E.M. (2007a). Retrograde neuronal tracing with a deletion-mutant rabies virus. Nature methods *4*, 47-49.

Wickersham, I.R., Lyon, D.C., Barnard, R.J.O., Mori, T., Finke, S., Conzelmann, K.-K., Young, J.a.T., and Callaway, E.M. (2007b). Monosynaptic restriction of transsynaptic tracing from single, genetically targeted neurons. Neuron *53*, 639-647.

Xue, Y., Ray, N., Hugh, J., and Bigras, G. (2016). Cell Counting by Regression Using Convolutional Neural Network. In Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I, G. Hua, and H. Jégou, eds. (Cham: Springer International Publishing), pp. 274-290.

Yamawaki, N., Suter, B.A., Wickersham, I.R., and Shepherd, G.M. (2016). Combining Optogenetics and Electrophysiology to Analyze Projection Neuron Circuits. Cold Spring Harb Protoc *2016*, pdb prot090084.

Yizhar, O., Fenno, L.E., Davidson, T.J., Mogri, M., and Deisseroth, K. (2011). Optogenetics in neural systems. Neuron *71*, 9-34.

Zeiler, M.D., Taylor, G.W., and Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. Paper presented at: 2011 International Conference on Computer Vision (IEEE).

Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision *73*, 213-238.