

Dissertation
submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by
M.Sc. Jeongbin Park

born in: Suwon, Republic of Korea

Oral examination: 21-04-2020

Segmentation-free inference of cell types
from *in situ* transcriptomics data

Referees: Prof. Dr. Roland Eils
Prof. Dr. Benedikt Brors

This work is licensed under a Creative Commons “Attribution-NonCommercial-NoDerivs 3.0 Unported” license.



Abstract

Recent advances in the fields of genome editing, whole-genome sequencing, single-cell RNA sequencing, and *in situ* spatial transcriptomics have enabled the cost-efficient production of high-throughput big data. However, the lack of dedicated bioinformatics algorithms to analyze such data has been a big hurdle. In this thesis, several novel bioinformatics tools applicable to each field are presented.

First, a series of web-based tools for genome editing are presented: Cpf1-Database, Cas-Analyzer, web-based Digenome-seq software, BE-Designer/Analyzer. These tools have been developed to guide researchers to easily use genome editing systems, using Cas9 or Cpf1, by providing an easily accessible web-based interface.

Second, the development of two bioinformatics pipelines are described: a small variant calling pipeline to process tumor genome sequencing data without a matched control, and a pipeline to pre-process single-cell RNA sequencing data.

Third, a novel segmentation-free algorithm to call cell-types from *in situ* transcriptomics data, namely Spot-based Spatial cell-type Analysis by Multidimensional mRNA density estimation (SSAM) is presented. Recent advances of *in situ* spatial transcriptomics techniques, such as multiplexed fluorescence *in situ* hybridization or *in situ*/intact tissue sequencing have enabled the discovery of spatial heterogeneity of cell types at the tissue level. However, cell type calling methods are often limited by cell segmentation algorithms due to various imaging problems. SSAM circumvents these problems by estimating spatial gene expressions as a density estimation of the mRNA in a spatial context and identifying *de novo* cell-types and their spatial organization without the need to segment cells. Optionally, SSAM can be guided by external sources of cell-type information, integrating them in a spatial context. In this thesis, SSAM is demonstrated with three different mouse brain tissues imaged by different imaging techniques: the somatosensory cortex (SSp) imaged by osmFISH; the hypothalamic preoptic region (POA) by MERFISH; and the visual cortex (VISp) by multiplexed smFISH. SSAM can produce similar results compared to segmentation-based methods and outperforms them when cell segmentation is the limiting factor.

In summary, the bioinformatics tools presented in this thesis overcome major obstacles that would normally hinder effective analysis: the web-based tools for genome editing have a wide

user base due to their easy-to-use web-based interfaces; omics data analysis pipeline that enables fast analysis of omics data utilizing a compute cluster and facilitate hypothesis generation when lacking control tissue; and SSAM that enables the analysis of *in situ* spatial transcriptomics data without being limited by cell segmentation. All of the tools and pipelines described in this thesis are open-sourced and freely accessible for non-profit, research-purpose use.

Zusammenfassung

Die jüngsten Fortschritte auf den Gebieten der Genomeditierung, der Gesamtgenomsequenzierung, der Einzelzell-RNA-Sequenzierung und der räumlichen *in-situ*-Transkriptomik haben die kosteneffiziente Produktion von großen Datenmengen - der sogenannten "big data" - mit hohem Durchsatz ermöglicht. Das Fehlen dedizierter Bioinformatik-Algorithmen zur Analyse solcher Daten war jedoch eine große Hürde. In dieser Arbeit werden verschiedene neuartige Bioinformatik-Tools vorgestellt, welche in jedem dieser Bereiche anwendbar sind.

Zunächst wird eine Reihe von webbasierten Werkzeugen zur Bearbeitung des Genoms vorgestellt: eine Cpf1-Database, Cas-Analyzer, webbasierte Digenome-seq-Software und BE-Designer/Analyzer. Diese Tools wurden entwickelt, um Forschern die einfache Anwendung von Genomeditiersystemen mithilfe von Cas9 oder Cpf1 zu erleichtern, indem eine leicht zugängliche, webbasierte Oberfläche bereitgestellt wird.

Zweitens wird die Entwicklung von zwei Bioinformatik-Pipelines beschrieben: zum einen eine Pipeline zur Identifikation kleiner Varianten in Tumorgenom-Sequenzierungsdaten ohne passender Kontrolle, zum anderen eine Pipeline zur Vorverarbeitung von Einzelzell-RNA-Sequenzierungsdaten.

Drittens wird ein neuartiger segmentierungsfreier Algorithmus namens "Spot-based Spatial cell-type Analysis by Multidimensional mRNA density estimation (SSAM)" vorgestellt, mit dem Zelltypen aus *in situ*-Transkriptomikdaten bestimmt werden können. Jüngste Fortschritte bei *in-situ*-Techniken zur räumlichen Transkriptomik, wie zum Beispiel Multiplex-Fluoreszenz-*in-situ*-Hybridisierung oder *in-situ*-Verfahren für die Sequenzierung intakter Gewebe, haben die Entdeckung räumlicher Heterogenität von Zelltypen auf Gewebeebene ermöglicht. Allerdings sind die Verfahren zur Bestimmung der verschiedenen Zelltypen in den auf Segmentierung beruhenden Algorithmen häufig aufgrund verschiedener Probleme bei der Bildaufnahme limitiert. SSAM umgeht diese Probleme, indem räumliche Genexpression als Dichteschätzung der mRNA in einem räumlichen Kontext geschätzt und *de-novo*-Zelltypen und ihre räumlichen Organisationen identifiziert werden, ohne dass Zellen segmentiert werden müssen. SSAM beinhaltet einen optionalen überwachten Lernalgorithmus der zelltypenbezogene Information in einem räumlichen Kontext integriert. In dieser Arbeit wird die Anwendung von SSAM anhand drei verschiedener Maushirngewebe demonstriert, die mit verschiedenen Bildaufnahmetechniken

niken abgebildet wurden: dem somatosensorischen Kortex (SSp), der mit osmFISH abgebildet wurde; die hypothalamische preoptische Region (POA) von MERFISH; und der visuelle Kortex (VISp), welcher durch multiplexiertes smFISH abgebildet wurde. SSAM kann im Vergleich zu den Ergebnissen der segmentierungsbasierten Methoden ähnliche Ergebnisse erzielen und diese übertreffen, wenn die Einschränkung durch die Zellensegmentierung der limitierende Faktor ist.

Zusammenfassend lässt sich sagen, dass die in dieser Arbeit vorgestellten Bioinformatik-Tools wichtige Hindernisse überwinden, die normalerweise eine effektive Analyse beeinträchtigen würden. Die webbasierten Tools für die Genombearbeitung haben aufgrund ihrer benutzerfreundlichen webbasierten Schnittstellen eine breite Anwenderbasis. Die Omics-Datenanalyse-Pipelines ermöglichen eine schnelle Analyse von Omics-Daten mithilfe eines Rechenclusters und erleichtern die Erstellung von Hypothesen, wenn kein Kontrollgewebe vorhanden ist. SSAM ermöglicht die Analyse von *in-situ* räumlichen Transkriptomikdaten, ohne durch Zellsegmentierung eingeschränkt zu sein. Alle in dieser Arbeit beschriebenen Tools und Pipelines sind Open-Source-Tools und für gemeinnützige Zwecke zu Forschungszwecken frei zugänglich.

Contents

Abstract	vii
Zusammenfassung	ix
List of Figures	xiii
List of Tables	xvi
1 Assessment of genome editing results	1
1.1 Introduction	1
1.1.1 Genome editing via CRISPR-Cas derived RGENs	1
1.1.2 Off-target effect of RGENs	5
1.1.3 Assessment of genome editing outcome by RGENs	5
1.1.4 Overview of the results	7
1.2 Results	7
1.2.1 Web-based database of Cpf1 target sites	7
1.2.2 Web-based assessment of genome editing sequencing results	12
1.2.3 Web-based assessment of off-target effects	14
1.2.4 Web-based design and assessment of CRISPR base editing	16
1.2.5 Analysis on the claim “unexpected mutations occurred by Cas9”	22
1.3 Discussion	24
1.4 Contributed publications	25
1.4.1 Work started previously and finished in Heidelberg	25
1.4.2 Work solely done in Heidelberg	25
2 Omics data analysis pipeline development	27
2.1 Introduction	27
2.1.1 Big omics data analysis using bioinformatics pipelines	27
2.1.2 Small variant calling without matched controls	29
2.1.3 Single cell RNA sequencing data analysis	30

2.2	Results	30
2.2.1	No control variant calling pipeline	30
2.2.2	Use case: No control variant calling for the ovarian cancer samples	31
2.2.3	Single cell RNA-seq data preprocessing pipeline	34
2.2.4	Use case: Pheno-seq project	34
2.3	Discussion	39
2.4	Contributed publications	39
3	Segmentation-free inference of cell types from <i>in situ</i> transcriptomics data	43
3.1	Introduction	43
3.2	Results	44
3.2.1	The SSAM computational framework	44
3.2.2	Analysis of mouse somatosensory cortex (SSp) imaged by osmFISH	49
3.2.3	Analysis of mouse hypothalamic preoptic area (POA) imaged by MER-FISH	69
3.2.4	Analysis of mouse visual cortex (VISp) imaged by multiplexed smFISH	90
3.3	Discussion	91
3.3.1	KDE bandwidth and lattice spacing	92
3.3.2	Possible extension of SSAM for <i>in situ</i> sequencing methods	99
3.3.3	Method details	99
3.4	Contributed publications	105
	Acknowledgements	125

List of Figures

1.1	The scheme of RGEN.	3
1.2	An example “special pattern” at a cleavage site of the Digenome-seq data.	6
1.3	The main page of the Cpf1-Database.	9
1.4	The quick info dialog.	10
1.5	The result page of Cpf1-Database.	11
1.6	Cas-Analyzer overview.	13
1.7	The main page of the web-based Digenome-seq data analysis tool.	17
1.8	The result page of an example Digenome-seq data analysis result.	18
1.9	The main page of BE-Designer.	19
1.10	The result page of BE-Designer.	21
1.11	The result of comparative analysis of mutations in FVB, F03, and F05 mice.	23
2.1	Ovarian cancer (OC) samples having numerous copy number variations (CNVs) but few small somatic variations.	32
2.2	Overview of single-cell RNA sequencing data preprocessing workflow.	33
2.3	NextSeq sequencing library structure of Wafergen data.	35
2.4	Quality control plots from one example sample provided by the preprocessing workflow.	36
2.5	Pheno-seq analysis of MCF10CA cell line.	37
2.6	Pheno-seq result of colorectal sample.	38
3.1	Schematic diagram of the SSAM computational workflow for cell type and tissue domain definition based on gene expression data.	45
3.2	Identification of intracellular regions via mRNA density estimation.	46
3.3	Generation of the cell-type map and the domain map.	48
3.4	SSAM improves astrocyte and ventricle detection in the mouse SSp region.	50
3.5	Local maxima selection and filtering criteria in the mouse SSp osmFISH dataset.	52
3.6	Cell-type signature identification and mapping in the mouse SSp osmFISH dataset.	53
3.7	SSAM identifies cortical layer tissue domains in the mouse SSp cortex.	55

3.8	Comparison of the cell type (Pericyte) and the corresponding marker gene expression (Vtn) of osmFISH data.	62
3.9	Comparison of the cell type (Vascular Smooth Muscle) and the corresponding marker gene expression (Acta2) of osmFISH data.	63
3.10	Comparison of the cell type (Inhibitory Vip) and the corresponding marker gene expression (Vip) of osmFISH data.	64
3.11	Comparison of the cell type (Inhibitory Pthlh) and the corresponding marker gene expression (Pthlh) of osmFISH data.	65
3.12	Comparison of the cell type (Perivascular Macrophages) and the corresponding marker gene expression (Mrc1) of osmFISH data.	66
3.13	Comparison of the cell type (C. Plexus) and the corresponding marker gene expression (Ttr) of osmFISH data.	67
3.14	Comparison of the cell type (Endothelial 1) and the corresponding marker gene expression (Apln) of osmFISH data.	68
3.15	SSAM 3D cell-type map confirms the rich diversity of heterogeneous cells in the posterior hypothalamic POA.	70
3.16	Local maxima selection, cell-type signature identification, and mapping in the mouse POA MERFISH dataset.	72
3.17	SSAM identifies enriched inhibitory and excitatory tissue domains in the posterior hypothalamic POA.	74
3.18	Side-by-side comparison of SSAM <i>de novo</i> cell-type map vs. cell segments presented by Moffit et al.	82
3.19	Side-by-side comparison of Astrocytes identified by SSAM guided mode vs gene expression of Aldh11l1, in MERFISH dataset.	83
3.20	SSAM identifies a new cell type in L4 and confirms rare Sst Chodl cell type in the mouse VISp region.	84
3.21	Local maxima selection and cell-type signature identification in the mouse VISp smFISH dataset.	85
3.22	Rescuing rare Sst Chodl cell types, and identifying new sub-layering in the L4 cortical layer in the mouse VISp smFISH dataset.	87
3.23	Rare Sst Chodl cell type localizes to the L5b cortical layer of the mouse VISp region.	89
3.24	Effect of KDE bandwidth and lattice spacing in the first region.	94
3.25	Effect of KDE bandwidth and lattice spacing in the second region.	95
3.26	Effect of KDE bandwidth and lattice spacing in the third region.	96
3.27	Effect of KDE bandwidth and lattice spacing in the fourth region.	97

3.28	Correlations between the <i>de novo</i> clusters with bandwidth 2.5 and the unmerged clusters with different bandwidths.	98
3.29	The preliminary cell-type map with mPFC imaged by STARmap	100

List of Tables

1.1	Various types of RGENs (RNA-guided endonucleases).	2
3.1	The matching score between the osmFISH PolyA segmented cells vs. SSAM cell-type map guided by the segmentation-based cell-type signatures of osmFISH data.	57
3.2	The matching score between the osmFISH PolyA segmented cells vs. SSAM cell-type map guided by the scRNA-seq cluster centroids for selected cell types based on high gene expression correlation.	58
3.3	The matching score between the osmFISH PolyA segmented cells vs. SSAM cell-type map guided by the scRNA-seq cluster centroids for selected cell types based on high gene expression correlation.	60
3.4	The matching score between the MERFISH segmented cells vs. SSAM cell-type map guided by the segmentation-based cell-type signature of MERFISH data.	78
3.5	The matching score between the MERFISH segmented cells vs. SSAM cell-type map guided by scRNA-seq cluster centroids for selected cell types based on high gene expression correlation.	79
3.6	The matching score between the MERFISH segmented cells vs. SSAM cell-type map guided by scRNA-seq cluster centroids for selected cell types based on high gene expression correlation.	81

Chapter 1

Assessment of genome editing results

1.1 Introduction

1.1.1 Genome editing via CRISPR-Cas derived RGENs

Short history of CRISPR-Cas genome editing

The CRISPR-Cas system is an immune system found in some archaea and bacteria, to protect themselves from the invasive foreign genetic elements [1, 2]. The system comprises of two parts, the CRISPR loci, and Cas (CRISPR associated) proteins. The CRISPR loci serve as acquired immune memory by containing short fragments of the invasive sequences, e.g. from phages, as spacers between specific repeat sequences. The crRNA (CRISPR RNA), which is the transcript of the spacer, is used as a template sequence that guides Cas proteins to bind to the target invasive sequence. After that, the invasive sequence is cleaved by the Cas protein guided by crRNA, so that the sequence cannot be used for the production of viral proteins.

Since the Cas proteins rely on crRNA-DNA hybridization for cleavage, it was suggested that the system can be used for genome editing by loading a desired RNA sequence to the Cas proteins to target any DNA sequence [3,4]. There are two classes of the CRISPR systems (class 1 and 2), among them the class 2 CRISPR system was first identified to have the potential to be used in genome editing because of the simplicity of the system. The class 2 CRISPR system only requires a single endonuclease rather than having to form a complex of several different Cas proteins, which is the case for the class 1 CRISPR system. The first CRISPR-Cas component discovered was the Cas9 endonuclease of *Streptococcus Pyogenes* (SpCas9). The endogenous SpCas9 dependent system requires 3 components - two short RNAs (crRNA and tracrRNA) and the SpCas9 endonuclease, but it was reported that the crRNA and tracrRNA RNAs can be fused to a single-chain guide RNA (sgRNA), which made the system much easier to use [3]. In January 2013, several different groups independently reported successful genome editing of a living organism using SpCas9 for the first time [5–10].

Type	Abbreviation	Organism	5'-PAM-3'	References
Cas9	SpCas9	<i>Streptococcus pyogenes</i>	NGG or NRG	[3]
	StCas9	<i>Streptococcus thermophilus</i>	NNAGAAW	[11]
	NmCas9	<i>Neisseria meningitidis</i>	NNNNGMTT	[12]
	SaCas9	<i>Staphylococcus aureus</i>	NNGRRT or NNNRRT	[13]
	CjCas9	<i>Campylobacter jejuni</i>	NNNVRYAC or NNNNRYAC	[14]
	SpaCas9	<i>Streptococcus pasteurianus</i>	NNGTGA	[13]
	Nme2Cas9	<i>Neisseria meningitidis</i>	NNNNCC	[15]
Engineered Cas9	VRER SpCas9		NGCG	[16]
	VQR SpCas9	<i>Streptococcus pyogenes</i>	NGA	[17]
	Xcas9 3.7		NGT or NG	[17]
Engineered Cas12b	BhCas12b v4	<i>Bacillus hisashii</i>	ATTN or DTTN	[18]
Cpf1	AsCpf1	<i>Acidaminococcus</i>	TTTN or TTTV	[19]
	LbCpf1	<i>Lachnospiraceae</i>		
	FnCpf1	<i>Francisella</i>	TTN or KYTV	[20]
Engineered Cpf1	RR AsCpf1	<i>Acidaminococcus</i>	TYCV	[21]
	RVR AsCpf1		TATV	

Table 1.1. Various types of RGENs. The list of reported RGENs with various different PAM sequences. This is not a full list of all RGENs currently available.

RNA-guided endonucleases (RGENs)

To date, many Cas9 variants [11–17], and also other RNA-guided endonucleases, such as Cpf1 (Cas12a) [19] or Cas12b [18], have been discovered (Table 1.1). Since such endonucleases can be programmed by guide RNAs, it was proposed to use the term ‘RNA-guided endonucleases’ (RGENs) instead of using the term ‘CRISPR’, to avoid confusion with the CRISPR-Cas immune system [22].

Although many of these endonucleases can be used for genome editing, SpCas9 (whose associated PAM sequence is 5'-NGG-3') is currently the most widely used thanks to the requirement of a short PAM sequence length, which naturally allows for a lot of target sites. However, although SpCas9 has many potential target sites in the whole genome of various organisms, there will be genes that cannot be targeted by SpCas9 due to the lack of the required PAM sequence in the genomic vicinity of these genes. Therefore, RGENs other than SpCas9, or engineered SpCas9 which recognize different PAM sequences, are additionally being used to target such genes.

Figure 1.1 shows the basic scheme of a typical RGEN (depicted here Cas9) and sgRNA complex. The sgRNA has a ‘spacer’ sequence, which is complementary to a so-called “proto-spacer’ sequence of the target DNA (Figure 1.1A). After the sgRNA is loaded into the RGEN (Figure 1.1B), the complex first recognizes the PAM sequence in the target DNA (the PAM

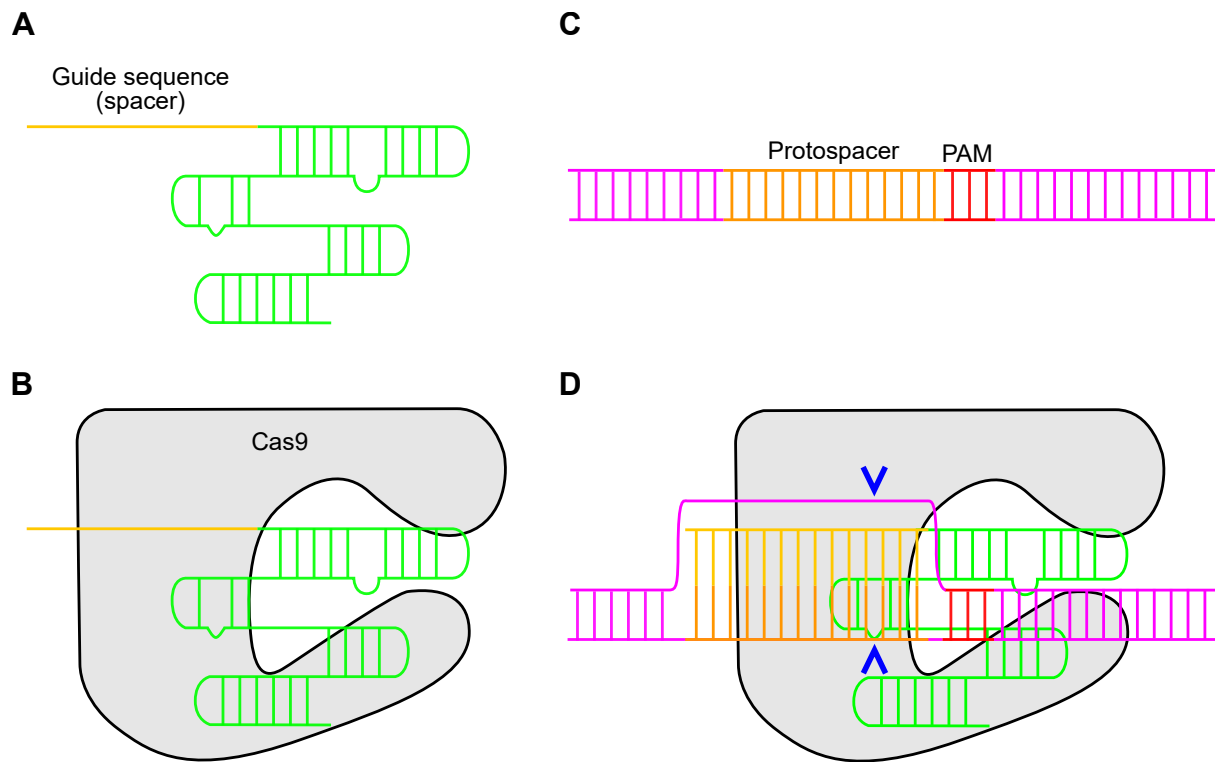


Figure 1.1. The scheme of RGEN.

Basic scheme of RGEN and sgRNA complex. This figure is based on the structure of Cas9 of *S. Pyogenes*, but other RGENs also have a similar structure. (A) Structure of single-chain guide RNA (sgRNA); (B) Cas9 and sgRNA complex; (C) Target DNA sequence; (D) Cas9 and sgRNA complex attached to target DNA. The cleavage positions by Cas9 are indicated with blue arrows.

sequence is a 2-6 base pair DNA motif used by Cas9 to bind to the target site (Figure 1.1C)). After the PAM sequence is recognized, the spacer sequence is hybridized to the protospacer sequence and the target DNA is cleaved, resulting in blunt-ended DNA (Figure 1.1D). If the protospacer and spacer are poorly hybridized, the RGENs cannot cleave the site.

Although Figure 1.1 is based on the structure of Cas9, other endonucleases, such as Cpf1 [19] and Cas12b [18], have a similar structure. The major difference is that 1) the PAM sequence of Cpf1 is located on the 5' side of the guide RNA, and 2) both Cpf1 and Cas12b make sticky ends instead of blunt ends after cleavage.

In addition to the capability of RGENs to make a double-strand break (DSB) at target sites, it was reported that it is possible to modify the Cas9 nuclease by introducing point mutations to the nuclease domain of Cas9. This has been done to make Cas9 nickase (shortly nCas9) which introduces a nick to DNA instead of DSB, or dead Cas9 (shortly dCas9) which do not have any cleavage activity but only binds to the target site. It is known that both nCas9 and dCas9 show the same specificity as unmodified Cas9 [4].

Genome editing with RGENs

Once RGENs have introduced DSBs at the target sites, genes can be knocked out via the endogenous non-homologous end joining (NHEJ) pathway which results in insertions or deletions (indels) near the DSB location during the repairing mechanism. For a successful gene knock-out, frameshift mutations are preferred since non-frameshift mutations can still result in the production of a functional protein. If there is a microhomology near the DSB site, DSBs can be repaired by an alternative NHEJ pathway, called the microhomology-mediated end joining (MMEJ) pathway. By using this, it is reported that non-frameshift mutations can be avoided by predicting mutation patterns based on microhomology which increases the efficiency of targeted gene knock-out [23].

Besides the removal of specific sequences, additional sequences can be inserted into the genome by taking advantage of the homology-directed repair (HDR) mechanism. This method works by transfecting the cells with an additional donor plasmid or short DNA fragment that contains the desired sequence as a template, called single-stranded donor oligodeoxynucleotide (ssODN), together with RGENs. Both ends of the ssODN have homology to the DNA sequences nearby the DSB site so that the ssODN can be used as a template sequence during HDR, which will eventually lead to the insertion of the desired sequence at the target site of the genome.

However, such genome editing approaches rely on DSBs that can lead to unwanted mutations, e.g. megabase-scale large indels [24] or mutations at off-target sites, which can lead to unwanted results. Therefore, new techniques that do not rely on DSBs have been developed, including CRISPR base editing. The CRISPR base editing uses a cytidine or a guanine deaminase, such as *APOBEC1* (apolipoprotein B editing complex 1) or *AID* (activation-induced deaminase)

linked to dCas9 [25–28]. As mentioned, dCas9 preserves the original Cas9 specificity but does not cleave the DNA. Therefore, it can be used to deliver the deaminase to the targeted region, which results in a highly specific DNA base substitutions near the target site.

1.1.2 Off-target effect of RGENs

As shown in Figure 1.1, the two requirements for RGENs to cleave the target site are 1) the existence of a PAM sequence, 2) the hybridization of guide RNA (spacer) and DNA (protospacer) sequences. However, it is reported that RGENs can cleave DNA even with a small number of mismatches between guide RNA and DNA sequences [29–31]. This means that RGENs can also bind and cleave ‘off-target’ sites which have protospacer sequences that differ from the guide RNA in several nucleotides, resulting in undesired mutations or chromosomal rearrangements. Therefore, in order to design a good guide RNA that does not have many potential off-targets, it is crucial to predict all such potential off-target sites in the whole genome of the organism to be edited. Many *in silico* approaches that were developed to design guide RNA of RGENs [32–40], including our fast GPU-powered algorithms – Cas-OFFinder [34] and Cas-Designer [35], thus incorporate the prediction of all potential off-targets in whole genomes allowing for several mismatches.

In addition to the *in silico* prediction of the off-targets, several experiment-based approaches have been suggested to avoid the off-target effect. One clever approach is using two Cas9 nickases (nCas9) to target two different sites in close genomic vicinity to each other [41, 42]. Introducing two nicks in the opposite strands of the targeted DNA region results in a sticky-ended double-strand break at the on-target site, whereas off-target sites will have only a single nick. Since a single nick on DNA has a much lower possibility to result in unwanted mutations than DSBs, such unwanted mutations can be avoided at the off-target sites. Another similar approach is using two dCas9 proteins fused with a FokI nuclease, which targets two very close sites in the genome. Since FokI only cleaves DNA when it forms a dimer with another FokI, the on-target site will result in cleavage by FokI dimer at the center of the two sites, but no cleavage at all at the off-target sites. Also, recently it has been reported that the specificity of Cas9 can be increased by engineering the Cas9 protein itself [16, 43].

1.1.3 Assessment of genome editing outcome by RGENs

Although several methods to increase the specificity of RGENs exists, it is still important to assess the outcome of genome editing.

Firstly, it is important to check the possible mutations at off-target sites. This can be achieved by several tools which have been developed so far, including our previous approach Digenome-seq [45–50]. Digenome-seq can detect insertions or deletions with the mutation rate below

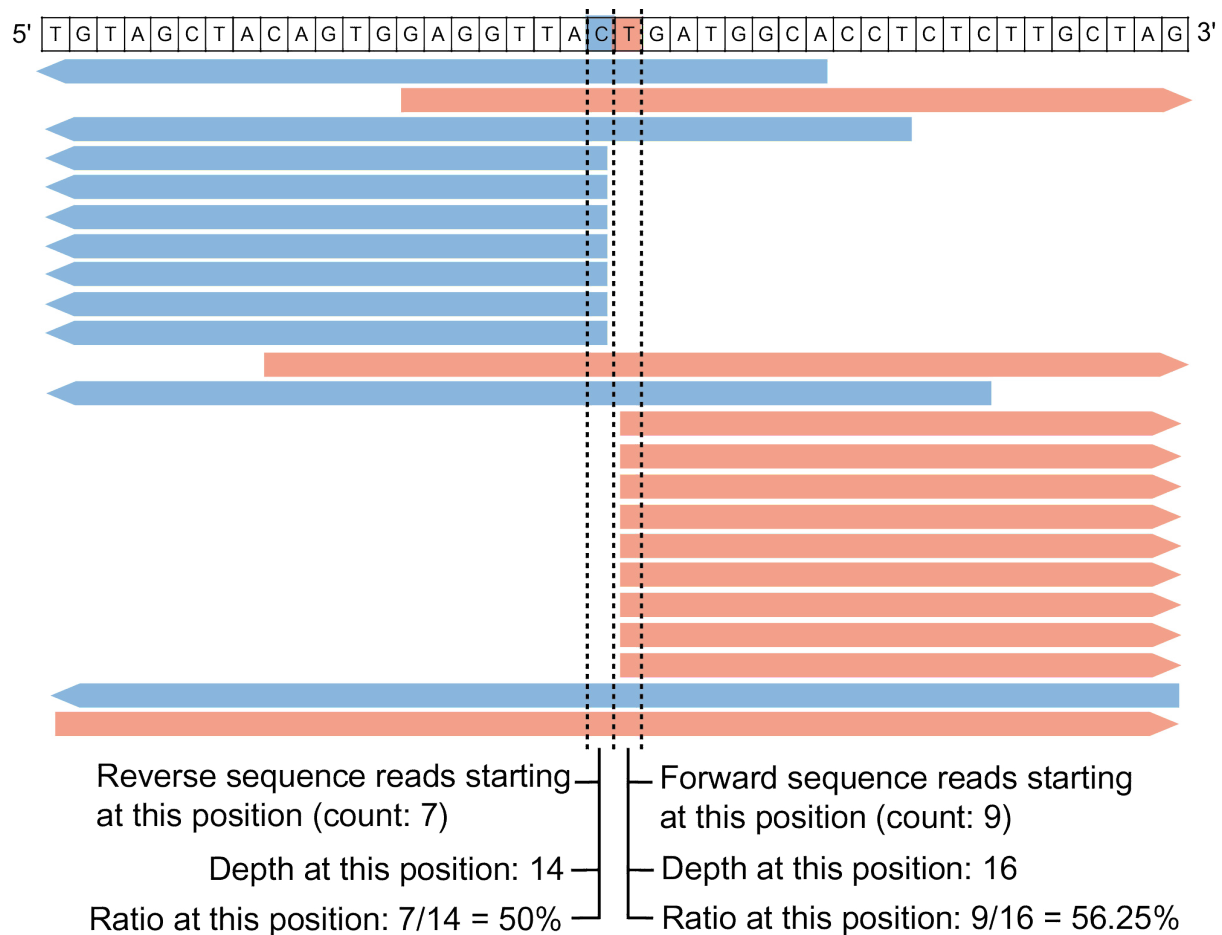


Figure 1.2. An example “special pattern” at a cleavage site of the Digenome-seq data.

(Note: This figure was jointly produced with Dr. Sangsu Bae, later published as a part of Park *et al.* [44].)

This figure shows an example special pattern of the same 5' reads starting at the same genomic location. At the 1 bp left and the right side of the cleavage site, the ratio of the number of reads starting at the position to the number of the sequencing depth is calculated and used to automatically identify the cleavage locations.

0.1 %, near the detection limit of targeted deep sequencing, and is, therefore, a sensitive and unbiased *in vitro* approach to profile genome-wide mutations at the off-target sites. The approach relies on *in vitro* cleavage of RGEN-transfected cells. The whole genomes of the RGEN-transfected cells are extracted and digested again with RGEN *in vitro* before library preparation, and then sequenced. Finally, the reads are mapped to the reference genome and visualized with a genomic viewer software (e.g. IGV). At an RGEN cleavage site, the *in vitro* cleavage made by RGEN results in special patterns of straight alignments (i.e., sets of 5' end of reads starts at the same genomic location) (Figure 1.2). These patterns can be also automatically detected by a computer program.

Secondly, it is also important to validate the editing result at each searched site in cells. This can be achieved by targeted deep sequencing near the searched site, which is known to be the most sensitive method due to its high precision. For DSB mediated gene knock-out and -in experiments, the insertion and deletion (indel) rate of different lengths caused by the DSB can be calculated. In addition, the knock-in rate can be calculated by screening for the presence of HDR template sequences near the DSB site. Moreover, it is also possible to identify the mutation rate of the base editing result with targeted deep sequencing data.

1.1.4 Overview of the results

In this chapter, I present several new computational tools for designing good guide RNAs to avoid off-target effects and assessment of the genome editing outcomes. Importantly, each tool is developed with an easy-to-use web interface, increasing its accessibility. First I introduce Cpf1-Database, a database of Cpf1 target sites in various organisms to help researchers to design good guide RNAs [51]. Second, Cas-Analyzer, a tool for assessment of mutation rates at the on-target site based on the next-generation sequencing data [52]. Third, a new Digenome-seq data assessment tool, a fast algorithm to detect off-target mutations [44]. Fourth, BE-Designer and BE-Analyzer, tools to design guide RNAs and assess on-target mutation rates of base editors [53]. In addition, I briefly discuss the recent claim about lots of unexpected mutations occurred by RGENs [54], which has never been observed using the tools mentioned in this chapter.

1.2 Results

1.2.1 Web-based database of Cpf1 target sites

For a high efficacy of genome editing, it is recommended to several criteria, including the number of potential off-target sites allowing several mismatches (to minimize off-target effect), the number of transcript variants simultaneously covered by the target (to completely knock-out the gene), the microhomology-associated out-of-frame score (to avoid in-frame mutations), the

relative position in the coding sequence (CDS) (for a complete gene disruption), excluding 4 thymidine nucleotides in tandem (often recognized as a terminal signal by RNA Pol III). Since it is time-consuming to calculate these metrics every time when designing a new guide RNA, it is much more efficient to have a database of pre-calculated values with all of the possible targets in various organisms, so that good guide RNAs can be designed quickly. In addition to the previous database of SpCas9 target sites [55], I made a new database called Cpf1-Database, which is a database of Cpf1 target sites in various organisms and is accessible through an easy-to-use web interface.

Target selection for AsCpf1/LbCpf1 endonucleases

Whole genomes of 12 different organisms, including *Homo sapiens* (GRCh38), *Rattus norvegicus* (Rnor 6.0), *Mus musculus* (GRCm38), *Danio rerio* (GRCz10), *Sus scrofa* (Ensembl v10.2), *Arabidopsis thaliana* (TAIR10), *Musa acuminata* (Banana), *Vitis vinifera* (IGGP 12X), *Solanum lycopersicum* (SL 2.5), *Glycine max* (JGI v1.0), *Drosophila melanogaster* (BDGP6) and *Caenorhabditis elegans* (WBcel235), were automatically downloaded from the Ensembl database (version 86) [56] via a custom Python script using the BioMart API. Next, all targets which have the PAM sequence of AsCpf1/LbCpf1 (5'-TTTN-3'), and have a cleavage position within any CDS region of the genomes, were identified. In this step, the additional information of the targets; 1) the GC content, 2) relative cleavage position in CDS, 3) the number of exons covered by the target, and 4) the microhomology-associated out-of-frame score, was also collected.

Searching for potential off-target sites

Using Cas-OFFinder [34], our previous OpenCL-based potential off-target searching algorithm, the potential off-target sites of all selected target sites were identified in the whole genome of each organism. Our defined potential off-target sites contained the PAM sequence 5'-TTTN-3' close to a sequence with high sequence similarity to any of the target sites, allowing mismatches up to 2 nucleotides. From the Cas-OFFinder result, the counts of potential off-targets allowing mismatches 0, 1, and 2 were calculated. This step was done in parallel using cluster computer Chundoong (<http://chundoong.snu.ac.kr>). The cleavage locations of potential off-targets were identified and classified into CDS (coding sequence), UTR (untranslated region), intron, or intergenic regions. All information obtained in this step was stored in a SQL database server.

Web interface

The main page of the Cpf1-Database shows the list of all genes of a selected organism (Figure 1.3), with functionality to search genes with keywords, by the gene name, Ensembl ID, or gene description, so that the desired genes can be easily selected.

CRISPR
RGEn Tools About Cas-OFFinder Microhomology Cas-Designer Database ▾ Cas-Analyzer Digenome-Seq ▾ Base Editing ▾ CRISPR-Sub

Cpf1-Database

Optimal selection of guide RNAs on the genome scale.

Cpf1-Database is a genome-wide gRNA library design tool for Cpf1. It contains **all available targets** of *Acidaminococcus sp.* Cpf1 (AsCpf1) or *Lachnospiraceae bacterium* Cpf1 (LbCpf1) that recognizes 5'-TTTN-3' PAM sequences **in all coding sequence (CDS) regions**. Users can select optimal target sequences from thousands of genes at once simply by changing the filtering conditions ([see tutorial here](#)). In addition, please see a [web application programming interface \(web API\)](#) for information about advanced use. Current version of genome assemblies are noted [here](#).

Citation info: Park J. and Bae S. Cpf1-Database: web-based genome-wide guide RNA library design for gene knockout screens using CRISPR-Cpf1. *Bioinformatics* **34**, 1077-1079 (2018)

Organism Type: Organism:

Category: Search:

Total 20317 genes are listed.

Symbol	Gene ID	Description	Biotype	Action
ZNF552	ENSG00000178935	zinc finger protein 552 [Source:HGNC Symbol;Acc:HGNC:26135]	Protein coding	Add to Collection Quick Info
CTD-2583A14.9	ENSG00000269476	Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:M0QZU9]	Protein coding	Add to Collection Quick Info
FRRS1	ENSG00000156869	ferric-chelate reductase 1 [Source:HGNC Symbol;Acc:HGNC:27622]	Protein coding	Add to Collection Quick Info
ZNF404	ENSG00000176222	zinc finger protein 404 [Source:HGNC Symbol;Acc:HGNC:19417]	Protein coding	Add to Collection Quick Info

Collection (0 genes) [Import genes from file...](#) [Select optimal crRNAs](#)

Figure 1.3. The main page of the Cpf1-Database.

The main page shows the list of genes. By clicking the ‘Add to Collection’ button to put a certain gene to the ‘Collection’ at the bottom of the page, and later guide RNAs of the selected genes can be designed in a batch. The possible target sites of a certain gene can be previewed with ‘Quick info’ button (see also Figure 1.4).

Filter

of mismatches (0, 1, 2) - - [\[?\]](#)

Most common exons coverage [\[?\]](#)

Out-of-frame score more than [\[?\]](#)

GC contents % - % [\[?\]](#)

Relative cleavage pos in CDS % - % [\[?\]](#)

Exclude 4 repeated thymidines (TTTT) [\[?\]](#)

Starting with 'G' at 5' end [\[?\]](#)

Starting with 'GG' at 5' end [\[?\]](#)

Starting with 'GA' at 5' end [\[?\]](#)

Gene

Symbol	Gene ID	Gene Name
ZNF404	ENSG00000176222	zinc finger protein 404 [Source:HGNC Symbol;Acc:HGNC:19417]

Transcripts (including CDS only)

Transcript ID	Biotype	Number of Targets	Number of Filtered Targets
ENST00000324394	Protein coding	113	28
ENST00000587539	Protein coding	113	28

Target sites (graphic)

Target sites (filtered)

No.	Transcript ID [?]	Target (5' to 3') [?]	Location [?]	Out-of-frame score [?]	Relative cleavage pos in CDS (%) [?]	Coverage [?]	GC contents [?]	# of off-targets (0, 1, 2) [?]
49	ENST00000324394	TTTACATTTGTAGGGTTTCACACCATG	chr19:43873391:+	75.76	47.95	2	43.48%	1, 0, 0
	ENST00000587539				48.04			
52	ENST00000324394	TTTTCTGATGCAGACACATATGTCGAT	chr19:43873423:+	73.39	46.01	2	43.48%	1, 0, 0
	ENST00000587539				46.11			
53	ENST00000324394	TTTCTGATGCAGACACATATGTCGATA	chr19:43873424:+	77.64	45.95	2	39.13%	1, 0, 0
	ENST00000587539				46.05			
60	ENST00000324394	TTTAGACGTCATTCTCACCTTACAGAA	chr19:43873517:-	62.98	41.73	2	43.48%	1, 0, 0

Close

Figure 1.4. The quick info dialog.

The quick info dialog shows the overview of possible targets of a gene. The list of targets can be filtered based on several useful criteria using ‘Filter’ panel on top of the page.

The screenshot displays the Cpfl-Database result page. At the top, a 'Filter' section allows users to adjust search criteria. The 'Total count of sgRNAs for each gene' is set to 3. The filter options include:

- # of mismatches (0, 1, 2) [1] [0] [0] [?]
- GC contents [20] % - [80] % [?]
- Starting with 'G' at 5' end [?]
- Most common exons coverage [?]
- Relative cleavage pos in CDS [5] % - [50] % [?]
- Starting with 'GG' at 5' end [?]
- Out-of-frame score more than [60] [?]
- Exclude 4 repeated thymidines (TTTT) [?]
- Starting with 'GA' at 5' end [?]

Below the filter, there are two view options: 'Card View' (selected) and 'Detail View'. The main content area shows four gene cards:

- TRBV18** of Homo sapiens (GRCh38/hg38) - Human (Yellow indicator):

Selected guide-RNAs (5' to 3')
TTTCATTGGGCTGCATCTCAGTCTTGC
TTTCAGACCTTCTCTGGGAGCTGCCG
N/A
- ZNF404** of Homo sapiens (GRCh38/hg38) - Human (Green indicator):

Selected guide-RNAs (5' to 3')
TTTCTGATGCAGACACATATGTCGATA
TTTATTAGACATCGAAAAATCCACACT
TTTTATTAGACATCGAAAAATCCACAC
- FMO1** of Homo sapiens (GRCh38/hg38) - Human (Yellow indicator):

Selected guide-RNAs (5' to 3')
TTTGAGAGGAGCGATGACCTTGGGGGG
N/A
- ARPC3** of Homo sapiens (GRCh38/hg38) - Human (Red indicator):

Selected guide-RNAs (5' to 3')
N/A
N/A

At the bottom of the page, there is a 'Done!' status on the left and two buttons: 'Remove genes with green indicators' (green) and 'Download filtered sequences...' (blue).

Figure 1.5. The result page of Cpfl-Database.

The result page shows the designed guide RNAs for the selected genes in the ‘shopping cart’. The indicator icon (in green, yellow, or red) on the left side of the gene name shows whether the specified number of guide RNAs are selected for the gene, where the green indicator means all guide RNAs, red means no guide RNAs, yellow means more than one but not all guide RNAs were selected. At the bottom of the page, it is possible to download the guide RNAs of the genes with green indicator and then remove them from the page, so that the guide RNAs of the remaining genes can be further designed with less strict filtering criteria.

Every gene has a ‘Quick Info’ button which opens a window that contains a short description of the gene, followed by a list of all target sites within the gene CDS (Figure 1.4). The window has a panel that contains several criteria that can be used to filter the targets – GC content, relative position in the CDS, the number of transcripts covered by the target, and the number of potential off-targets per mismatched nucleotides.

Alternatively, the Cpf1-Database offers a way to design guide RNAs of multiple genes at once. By clicking the ‘Add to Collection’ button of each gene, the gene is collected in the bottom panel of the main page, similar to the ‘shopping cart’ functionality of shopping websites. After collecting the desired genes, by clicking the ‘Select optimal gRNAs’ button, the result page is shown with up to 3 (default) optimal guide RNAs per each selected gene (Figure 1.5), where the number of optimal guide RNAs can be adjusted in the result page. The result page also has a functionality to filter guide RNAs based on the same criteria in the ‘quick info’ window, which can be used to filter guide RNAs of all selected genes at once. Each gene has an indicator colored in red, yellow, and green, which stands for no guide RNAs, more than 0, and all optimal guide RNAs are designed, respectively. Finally, the result page has the functionality to remove the genes with green indicator – so that the less optimal guide RNAs of the remaining genes can be selected again with lowering the filtering criteria. By repeating this step, the optimal guide RNAs of the selected genes can be easily designed with only a few clicks.

In addition to the interactive web interface, the Cpf1-Database also offers a programmatic way to retrieve information via a web application programming interface (web API). The specification of the web API can be found at ‘http://www.rgenome.net/static/cpf1-database-help/web_api.pdf’.

1.2.2 Web-based assessment of genome editing sequencing results

After editing the genome of target cells with the selected optimal guide RNAs, the result of genome editing can be assessed by targeted deep sequencing near the on-target site. But analysis of such sequencing data usually requires multiple command-line software tools, not easily accessible to many researchers who are not familiar with the command-line interface. Therefore, I developed an easy-to-use web-based assessment tool, Cas-Analyzer. The main strength of Cas-Analyzer is that its core algorithm is purely implemented in JavaScript, therefore the data do not have to be uploaded to the remote server – which can avoid long data upload times and data security/privacy issues.

Web interface of Cas-Analyzer

Cas-Analyzer accepts input data and parameters via a web form. Cas-Analyzer requires paired-end or single-end deep sequencing data (Figure 1.6A), with the original reference sequence and



Figure 1.6. Cas-Analyzer overview.

(A) The input of Cas-Analyzer, which is a single or paired deep sequencing FASTQ files (compressed or decompressed); (B) The parameters for the Cas-Analyzer, the reference sequence, RGEN information, and optionally ssODN information for the gene knock-in rate analysis; (C) A schematic diagram shows how the parameters R and r works. R is used to define indicator sequences to filter valid sequences, and r is used to determine WT sequence; (D) The summary table, which shows the number of sequences in each category (insertion, deletion, and WT); (E) The interactive plot, which shows the location of indels and their size histograms; (F) the list of sequences, aligned to the reference sequence for an easy comparison.

the RGEN information (the type of RGEN and the target sequence) to identify the cleavage locations (Figure 1.6B). Optionally, the template HDR sequence can also be given such that it can be used to calculate the expected knock-in rate. The input sequencing data can be either raw FASTQ or gzip (or blocked gzip) compressed. In the case of compressed input data, a JavaScript library *pako* (<http://nodeca.github.io/pako/>) is used to decompress the input file before analysis (here, *pako* was slightly modified to be compatible to the blocked gzip files). For the paired-end data, *Fastq-join*, a part of *ea-utils* (<https://code.google.com/archive/p/ea-utils/>), was reimplemented in JavaScript and used by Cas-Analyzer.

Data analysis

Cas-Analyzer works in 3 steps to calculate mutation frequencies. First, Cas-Analyzer identifies the location of cleavage based on the input RGEN information and the reference sequence. The indicator sequences, i.e. R nucleotides on both sides, including the 12 nt left and right indicator sequences at both outer sides, are used to select valid sequenced reads, by selecting the reads with both indicator sequences allowing up to 1 mismatch in the sequencing data (Figure 1.6C). Second, the unique sequenced reads among all reads are selected and their frequencies are counted, and the reads with a frequency below n are excluded from the analysis. Third, the unique reads are classified into the following three groups: ‘insertion’, ‘deletion’, and ‘WT or substitution’, and each frequency is reported as a table (Figure 1.6D). Here, the parameter r defines “WT marker” sequence, defined as a short sequence in the reference sequence at the cleavage location (Figure 1.6C). Cas-Analyzer firstly finds the WT marker in the unique sequences, and classifies into ‘WT or substitution’ group if the sequences contain the WT marker; otherwise, it is classified into ‘insertion’ if longer than, or ‘deletion’ if shorter than the reference sequence. Optionally, if the HDR template sequence is given, the Cas-Analyzer defines the difference of the template sequence and the reference sequence as an HDR indicator and then uses this HDR indicator to classify the unique sequences into an additional ‘HDR’ category.

After the completion of the data analysis, the unique sequences are aligned to the reference sequence. For this, the EMBOSS Needle was reimplemented in JavaScript. The aligned unique sequences are presented in the result page with the classification, sorted by their frequency in descending order (Figure 1.6F). For user convenience, the result is also presented as interactive graphs on the result page (Figure 1.6E).

1.2.3 Web-based assessment of off-target effects

Although Digenome-seq can be used to detect off-target cleavages with high sensitivity [50], the analysis pipeline presented together with the Digenome-seq publication requires extensive command-line interactions and produces several large intermediate files, resulting in large space

requirements and a long-running time. Here, I present a redesigned web-based analysis tool for Digenome-seq data that solely runs on web browsers, capable of running instant analysis and presenting interactive result representation.

Implementation

The core algorithm of the web-based Digenome-seq data analysis tool was implemented in C++. HTSlib, which is a *de facto* standard library to process huge sequencing data, is used to read BAM files [57]. The reads with a mapping quality of 0 are excluded from the analysis by default. After retrieving the sequences, the algorithm detects DSB sites by finding a series of forward/reverse reads starting at the same position. By default, the tool requires a minimum value of 5 reads having a start/end at the same position. Also, it calculates the sequencing depth at the position and divides the number of reads having the same 5' position by the sequencing depth, to get the ratio of the reads at each position. By default, the threshold for filtering is set as 10% and 20 % for the sequencing depth and the ratio, respectively. The 5' sticky ends are shown as overlapping reads because the 5' sticky ends are filled with bases during the sequencing process. On the other hand, 3' sticky ends are removed by exonuclease activity before sequencing and result in a gap. Based on this, the cleavage score at the position i (S_i) is computed using the following formula, which is a generalized version of the one that was suggested before [58].

$$S_i = \sum_{a=1}^5 \frac{F_i - 1}{D_i} \times \frac{R_{i-4+G+a} - 1}{D_{i-4+G+a}} \times (F_i + R_{i-4+G+a} - 2) + \sum_{a=1}^5 \frac{R_{i-1+G} - 1}{D_{i-1+G}} \times \frac{F_{i-3+a} - 1}{D_{i-3+a}} \times (R_{i-1+G} + F_{i-3+a} - 2) \quad (1.1)$$

where:

F_i : Number of forward sequence reads starting at position i

R_i : Number of reverse sequence reads starting at position i

D_i : Sequencing depth at position i

G : Size of sticky end overhang (positive/negative value for 5'/3' overhang)

To build a web tool, HTSlib was firstly ported to work with the compiler Emscripten (<http://emscripten.org/>), which generates asm.js (a pre-optimized subset of JavaScript, <http://asmjs.org/>) or WebAssembly (<https://webassembly.org/>) with the C/C++/Rust written code, and used as a library in the core C++ code to read BAM files. The core C++ code, alongside the ported HTSlib, was compiled to asm.js with Emscripten. Therefore, all code for analysis conforms to the JavaScript standard, so it works well in any modern web browser, so analysis can be easily performed using a web browser without installing any local tools. The

content of the BAM files is not uploaded to any server, but streamed on the client-side, to cope with the memory limitation of web browsers and additionally achieve high performance. Since the data are not uploaded to the server, it can be analyzed immediately without preparation time, and the analysis via this tool is free from data security issues. The core algorithm was optimized by the Emscripten compiler, which allows for fast analysis in modern web browsers. In the benchmark, a full analysis of a 100GB BAM file took three hours to do a full analysis using the Intel i5 3570k central processing unit (CPU) on a single thread.

In addition to the web version of the analysis tool, the command-line version of the tool was also generated by compiling the same C++ code with the GNU C Compiler (GCC). The use of the command-line version of the tool allows users to run the analysis with even faster speed, or integrate the tool with their pipeline.

Web interface

As input, the tool takes a position-sorted BAM file containing the aligned sequences from a nuclease-treated sample, and optionally a control BAM file for comparison. The user has to specify a minimal set of required parameters, which include the cleavage type (blunt or sticky ended), minimum mapping quality, minimum number of reads with same 3' and 5' ends, minimum sequencing depth, minimum ratio of reads starting at the same location to the sequencing depth, and the minimum cleavage score described in equation 1.1 (Figure 1.7). Optionally, if the user supplies target protospacer sequences for the nucleases, the tool retrieves flanking sequences nearby the target site from Ensembl via its REST API (<https://rest.ensembl.org/> and <https://rest.ensemblgenomes.org/>). The protospacer sequence is then semi-globally aligned to the retrieved flanking sequence, using a custom-made alignment algorithm written in JavaScript (here, Dr. Liam Childs implemented the JavaScript semi-global alignment algorithm) (Figure 1.8). The result page shows the list of cleavage positions, together with interactive summary plots of the cleavage score versus the position in Manhattan/Circos plots, and optionally the alignment results. The interactive plots are generated by using D3.js (<http://d3js.org/>). The web tool is freely available at <http://www.rgenome.net/digenome-js>.

1.2.4 Web-based design and assessment of CRISPR base editing

Building on the previously described tool Cas-Designer, I made a dedicated guide RNA designing tool for CRISPR base editors, BE-Designer. Besides, I co-supervised a student (Mr. Gue-Ho Hwang) with Prof. Sangsu Bae at Hanyang University, Korea, to create a dedicated assessment tool for CRISPR base editors based on Cas-Analyzer, called BE-Analyzer.

Sequencing Data

Please select a nuclease-treated and a control BAM file (which contains mapped reads against reference genome). The BAM files need to be coordinate-sorted.

Nuclease-treated BAM file

파일 선택 선택된 파일 없음

Control BAM file (optional) [?](#)

파일 선택 선택된 파일 없음

Reference genome (retrieved from [Ensembl](#)):

Human (GRCh38) ▼

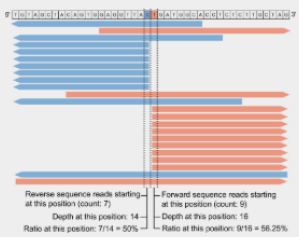
Nuclease Information

Cleavage type

Blunt end ▼

e.g.) Cas9: blunt end, AsCpf1, LbCpf1: 3 overhang (5'), FnCpf1: 5 overhang (5'), ZFN: 4 overhang (5'), TALEN: 6 overhang (5')

Target sequence(s), one sequence per line (optional, 5' to 3'):



Filtering Options

Minimum mapping quality for bam reads

Minimum number of forward reads with same 5' ends

Minimum number of reverse reads with same 3' ends

Minimum depth at each position

Minimum ratio at each position

Minimum cleavage score [?](#)

Run digenome-seq

Figure 1.7. The main page of the web-based Digenome-seq data analysis tool.

This figure shows the input of the web-based digenome-seq analysis tool, a nuclease-treated BAM file, and an optional control BAM file, reference genome, cleavage information, optional RGEN guide RNA sequences, and the filtering options.

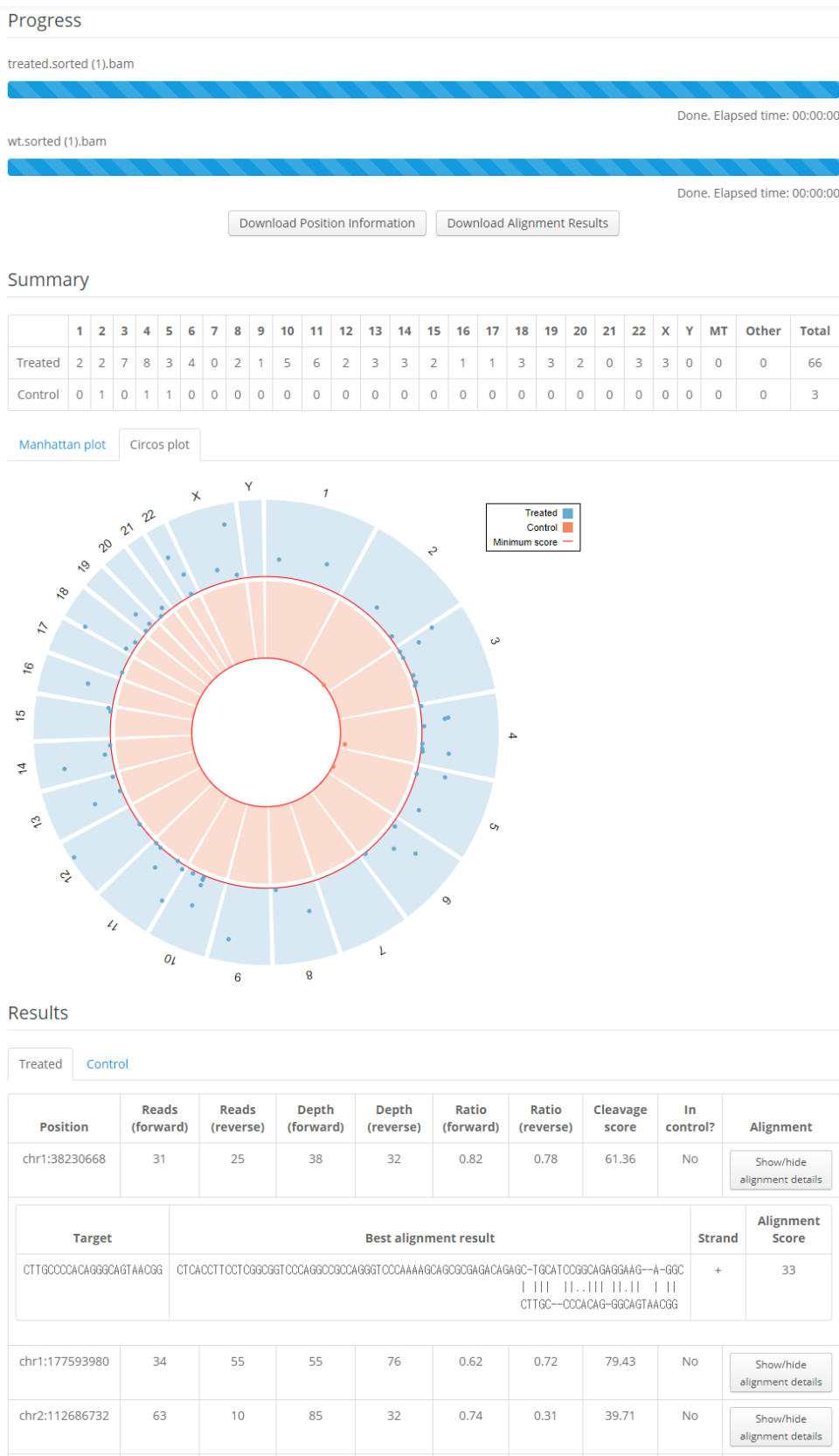


Figure 1.8. The result page of an example Digenome-seq data analysis result. This figure shows an example output of the digenome-seq web tool. The summary table shows the number of cleavage sites in each chromosome, and also presented as an interactive plot (manhattan or circos plot). The location of the cleavage sites are listed in bottom.

PAM Type

CRISPR-Cas orthologues for base editing

- SpCas9 from *Streptococcus pyogenes*: 5'-NGG-3'
- SpCas9-VQR from *Streptococcus pyogenes*: 5'-NGAN-3'
- SpCas9-EQR from *Streptococcus pyogenes*: 5'-NGAG-3'
- SpCas9-VRER from *Streptococcus pyogenes*: 5'-NGCG-3'
- SaCas9 from *Staphylococcus aureus*: 5'-NNGRRT-3'
- SaCas9-KKH from *Staphylococcus aureus*: 5'-NNNRRT-3'
- StCas9 from *Streptococcus thermophilus*: 5'-NNAGAAW-3' (W = A or T)
- CjCas9 from *Campylobacter jejuni*: 5'-NNNRYAC-3' (V = G or C or A, R = A or G, Y = C or T)
- xCas9 3.7 (TLIKDIV SpCas9) from *Streptococcus pyogenes*: 5'-NGT-3'
- AsCpf1 from *Acidaminococcus* or *LbCpf1* from *Lachnospiraceae*: 5'-TTTV-3' (V = G or C or A)
- AsCpf1 from *Acidaminococcus* or *LbCpf1* from *Lachnospiraceae*: 5'-TTTN-3'
- Spy-macCas9 from *Streptococcus pyogenes* and *Streptococcus macacae*: 5'-NAAN-3'
- Nme2Cas9 from *Neisseria meningitidis*: 5'-NNNCC-3'

Target Sequence

Insert any sequence(s) where you want to search for CRISPR base editing targets (raw sequence or FASTA format, maximum 1000 chars):

```
>homo sapiens FANCM, exon 2
GGTCTACACAAGCTTCCACCAGGAAGGAAATA
TGGTGCAAGTAAGAGAGTGCTTTTCTTACACC
TCAGGTCATGGTAAATGACCTTTCTAGAGGAG
CTTGCCCGCTGCTGAAATAAAGTGTTAGTT
ATTGATGAAGCTCATAAAGCTCTCGGAAACTA
TGCTATTGCCAG
```

crRNA length (length of target without PAM)

Or, upload a FASTA-formatted file containing target sequence(s) (maximum 1 KiB):

선택된 파일 없음

Base editing type:

Base editing window to

Target Genome

Organism Type

Genomes

- Homo sapiens (GRCh38/hg38) - Human
- Homo sapiens (hg19) - Human
- Mus musculus (mm10) - Mouse
- Bos taurus (bosTau7) - Cow
- Canis familiaris (canFam3) - Dog
- Rattus norvegicus (rn5) - Rat
-
-
-

Figure 1.9. The main page of BE-Designer.

This figure shows the input parameters of the BE-Designer, the type of PAM, the target genome, the target sequence where the guide RNAs will be found within, and the type of base editor.

BE-Designer

The web interface of BE-Designer is essentially the same as Cas-Designer [35], based on the Bootstrap library and Django web framework (Figure 1.9). The result page shows all of the information that Cas-Designer provides – e.g. the relative location of the target, GC content, and potential off-target sites. Additionally, BE-Designer shows the possible base edits within the target window near the target site with the information.

BE-Designer currently supports 4 different CRISPR base editors [25–28], based on the inactive form of one of the following RGENs: SpCas9 (5'-NGG-3') [as well as its variants: SpCas9-VQR (5'-NGAN-3'), SpCas9-EQR (5'-NGAG-3'), SpCas9-VRER (5'-NGCG-3'), xCas9 3.7 (TLIKDIV SpCas9; 5'-NGR-3' and 5'-NG-3')], StCas9 from *Streptococcus thermophilus* (5'-NNAGAAW-3'), CjCas9 from *Campylobacter jejuni* (5'-NNNVRYAC-3'), SaCas9 from *Staphylococcus aureus* (5'-NNGRRT-3') and its engineered variant, SaCas9-KKH (5'-NNNRRT-3'), AsCpf1 from *Acidaminococcus* or LbCpf1 from *Lachnospiraceae* (5'-TTTV-3' or 5'-TTTN-3'), Spy-macCas9 from *Streptococcus pyogenes* and *Streptococcus macacae* (5'-NAAN-3'), and Nme2Cas9 from *Neisseria meningitidis* (5'-NNNNCC-3').

BE-designer accepts the type of base editor, type of RGEN, and the desired DNA sequence in IUPAC nucleotide codes (with mixed bases) to be edited (as a raw text in the web form or a FASTA file), and the whole genome for potential off-target identification in the main page (Figure). After submitting the inputs, BE-Designer shows all possible target sites of the given RGEN in the DNA sequences as a list in the result page. In addition to the basic information, e.g. the relative location in the DNA sequence and the GC contents, both the nucleotides which can be edited by the selected base editor and the edited result (nucleotides and amino acids) are shown. Next, Cas-OFFinder is used to find the potential off-targets of the identified targets allowing up to 2 mismatches in the given whole-genome and shows the possible base edited results at the potential off-target sites (Figure 1.10).

The result page of BE-Designer is based on AJAX (Asynchronous JavaScript and Extensible Markup Language), so that the result is updated in real-time without the need of refreshing the whole result page. Based on this, BE-Designer has the functionality to instantly filter the results according to the GC content and the number of off-targets allowing up to 2 mismatches. For every off-target, a link to the Ensembl genome browser which shows the sequences, transcripts, and genes near the location of the off-target is provided, so that the more detailed information of the off-target can be easily accessed.

BE-Analyzer

(Note: BE-Analyzer was mainly developed by Mr. Gue-Ho Hwang. I only supervised his work, here I briefly introduce the rationale of the tool.)

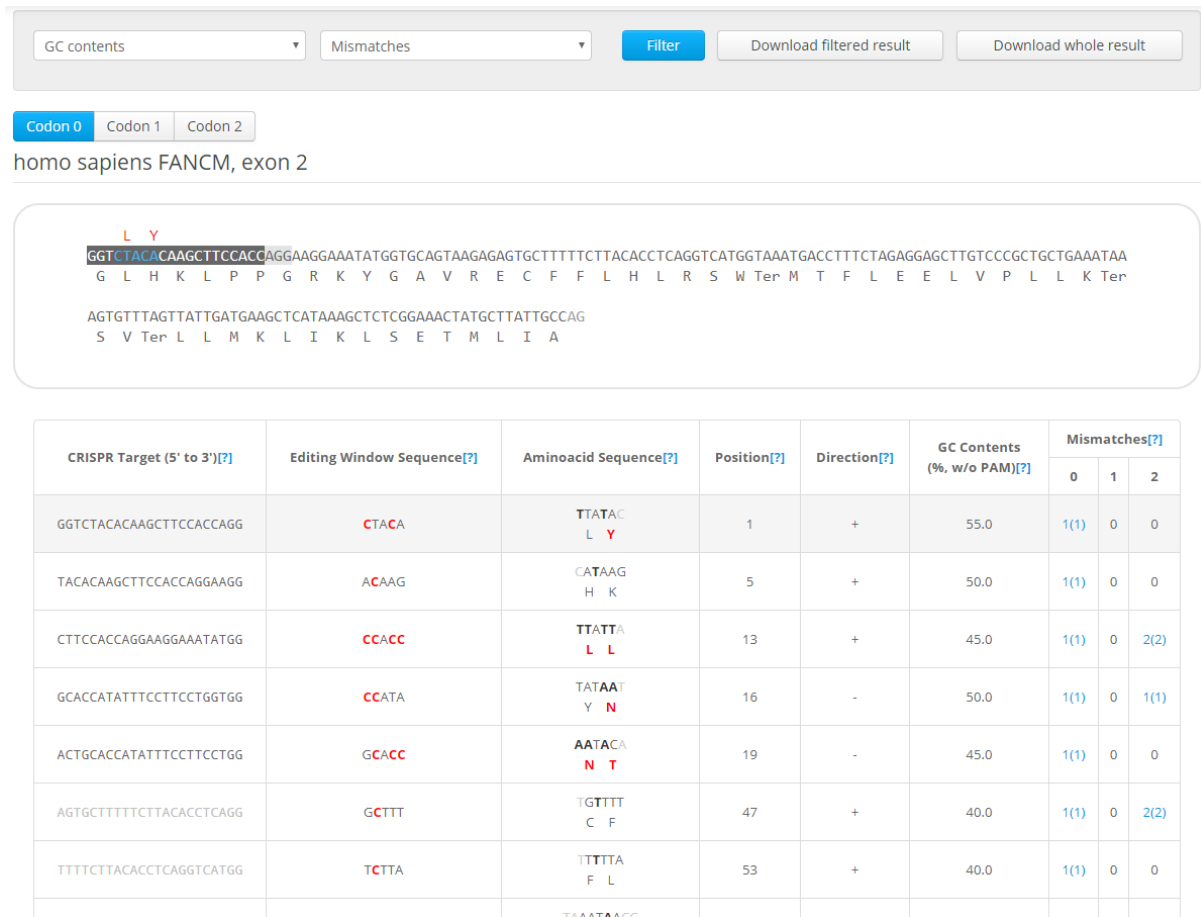


Figure 1.10. The result page of BE-Designer.

In the result page, a list of the designed guide RNAs are shown. The list also shows possible base editing results, both nucleotide and amino acid changes.

BE-Analyzer is a web-based assessment tool for CRISPR base editing results. The tool is largely based on Cas-Analyzer; The tool accepts the targeted deep sequencing data of a base editing result with several required parameters and produces an assessment result with almost the same interface and functionality as Cas-Analyzer. The main difference of the tool is that the tool mainly reports the mutation rate of the base changes, and additionally shows the according to amino acid changes. BE-Analyzer was published together with BE-Designer as a single article. Both tools are available at <http://rgenome.net/>.

1.2.5 Analysis on the claim “unexpected mutations occurred by Cas9”

In 2017, Schaefer *et al.* [54] reported that 1,397 unexpected single-nucleotide variants (SNVs) and 117 small insertions and deletions (indels) were found in two gene-edited mice by Cas9 (namely F03, F05) when it is compared to a wild-type control mouse (FVB) or a mouse genomic variation database.

However, most of the reported variation sites did not show sequence homology between the on-target site, nor protospacer-adjacent motif (PAM) sequence, which did not match with the previous reports: (1) It have been reported that Cas9 do not show detectable off-target effects when there are just two or three mismatches between the protospacer and guide RNA sequences [5, 29]; (2) Digenome-seq experiment did not find any unusual off-target sites in edited human genome [50, 58]. Moreover, there have been reports based on WGS data, which showed that it is rare for Cas9 to induce off-target indels in clonal cells or a genome-edited animal, which also contradicts to the result of Schaefer *et al.* [54] Also, the top 50 most likely potential off-target sites with 3 or 4 mismatches did not have any off-target mutations in the data, although the indel sites reported by Schaefer *et al.* [54] had mismatches with more than 10 nucleotides. Also, Schaefer *et al.* [54] reported that so many SNVs (1,397 SNVs over 117 indels), which is quite unlikely since the mutations induced by Cas9 is by NHEJ pathway after cleavage at the target site. Although it is known that SNVs can be made by the process, by removal of 1 bp and insertion of 1 bp, such event is very rare – at the rate of 1 % even at the on-target site, even less at the off-target sites. Overall, it is more likely that the SNVs and indels were not made by cleavages induced by Cas9.

Instead, one possible hypothesis is that the two gene-edited mice are genetically closer to each other when they are compared with the control mouse. In other words, the SNVs found in the two gene-edited mice were inherited by parents of the two mice, not induced by Cas9. This was discussed in the following two aspects.

First, Strelka [60] and Mutect [61], which were also used by Shaefer *et al.* [54], were used to call SNVs to test this hypothesis. The pairwise comparison of common SNVs between three mice revealed that the number of sample-specific SNVs of both F03 and F05 mice when it is

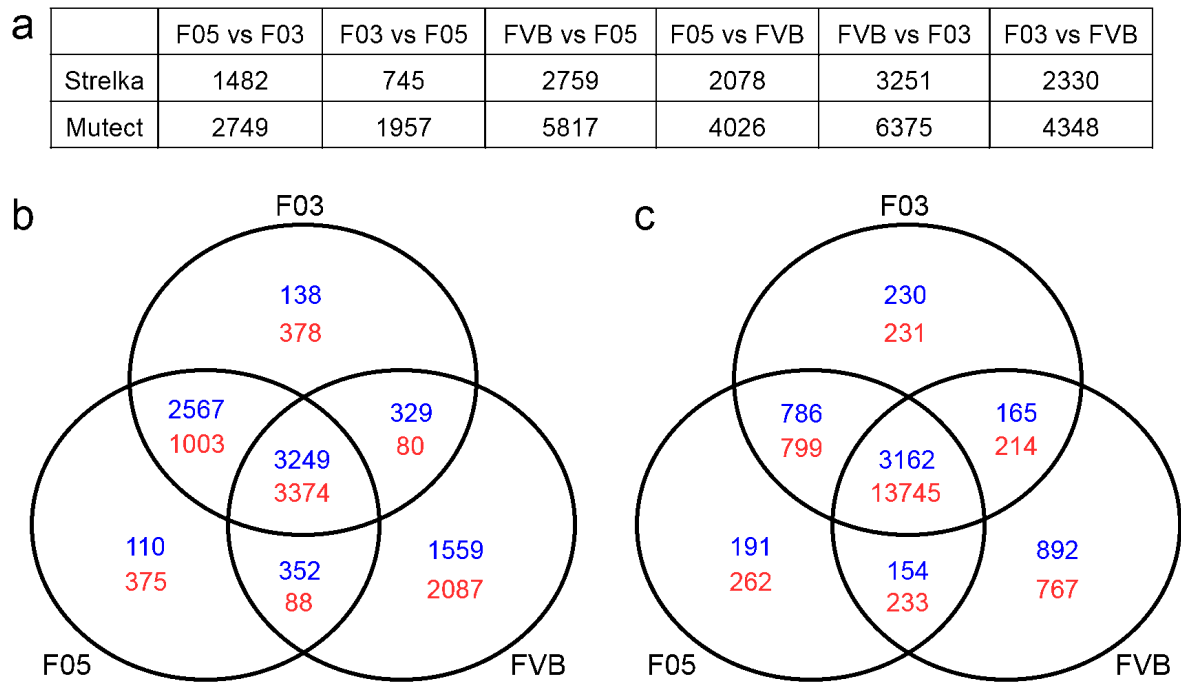


Figure 1.11. The result of comparative analysis of mutations in FVB, F03, and F05 mice. (Note: This figure has been published as a part of Kim *et al.* [59]. The figure was originally produced by myself.)

(a) pairwise comparison of the number of SNVs, by Strelka and Mutect; (b) The number of homozygous variants depicted in Venn diagram, the number of SNVs is denoted in blue color, and that of indels is in red color; (c) The number of heterozygous variants depicted in Venn diagram, as like panel (b).

compared to each other is considerably lower than the number of sample-specific SNVs of FVB mouse compared to F03 or F05 (Figure 1.11a).

Second, Platypus [62], a multi-sample variant caller, was used to call the SNVs and indels of the three mice altogether, to rule out a calling bias by the different sequencing depth of the three mice (Figure 1.11b,c). The analysis revealed two observations: 1) the number of unique homozygous variants in FVB mouse was much higher than that of unique variants in F03 and F05 mice, and 2) the number of shared homozygous variants between F03 and F05 mice was much higher than that of homozygous variants shared between F03 and FVB, or F05 and FVB mice. Whereas the number of unique heterozygous variants in all three mice was much smaller than that of shared heterozygous variants between all three mice, and the number of unique heterozygous variants was comparable to each other. Since mutagenic processes have higher chances to induce heterozygous variants and unlikely to make such homozygous variants, the comparable number of unique heterozygous variants in all three mice suggests that the variants are not mainly induced by unexpected mutagenic processes by Cas9. Instead, it makes more sense to conclude that the excess number of variants found in FVB mouse was mainly due to the close genetic distance between F03 and F05 mice, which was confirmed by the high number of shared homozygous variants between F03 and F05 mice.

1.3 Discussion

For a successful genome editing, it is important to design good guide RNAs by considering the number of off-targets, including those with several mismatches, as a badly designed RGEN can result in detrimental off-target effects. In addition to this, it is also important to assess on-target mutation rates, and also the off-target effects incurred by RGENs after the genome editing. In this thesis, I presented a series of computational tools to aid users to easily perform such tasks with a user-friendly web interface, minimizing required background knowledge of bioinformatics.

Cpfl-Database enables users to design thousands of guide RNAs of an organism with only a few clicks through an easy-to-use web interface, making the tool especially useful for genome-wide screening experiments using Cpfl. Cas-Analyzer and web-based Digenome-seq analysis tools can be used to easily assess genome editing outcomes. Cas-Analyzer is a useful tool to perform a high-resolution assessment of on-target mutation rates. The web-based Digenome-seq analysis tool provides an easy-to-use interface for the analysis of Digenome-seq data. Thanks to the totally redesigned, fully optimized algorithm of the tool, Digenome-seq data can be easily and quickly analyzed on web browsers, compared to the old pipeline which was published earlier. Both BE-Designer and BE-Analyzer are useful for CRISPR base editor experiments. BE-Designer is designed to aid users to select good guide RNAs with less potential off-targets

with taking care of the potential base changes, and BE-Analyzer enables analysis of the base changes and possible other mutations at the on-target position.

The efforts to provide a set of *in silico* tools are to establish a framework to minimize off-target mutations by RGENs. In that regard, I additionally discussed the recent claim about lots of unexpected mutation caused by RGENs [54], mainly why it is illogical, and about the investigations of possible mistakes that the authors made [59]. It was turned out that it is unlikely Cas9 caused the unexpected mutations in both F03 and F05 mice, but these are much better explained by the genetic background of the three mice. Therefore, it is difficult to conclude that Cas9 can cause “unexpected mutations” which cannot be avoided by any of the computational tools described in this chapter.

Overall, I presented a set of computational tools that can be used to design good guide RNAs to avoid off-target effects and assess the outcome of genome editing. Thanks to the easy-to-use web-based interface of the tools, anyone who is not familiar with the complicated command-line interface can instantly run analysis without a bioinformatics background. All of the tools are publicly available at <http://rgenome.net>.

1.4 Contributed publications

1.4.1 Work started previously and finished in Heidelberg

(Note: below publications have indications that the “current address” is in Heidelberg)

- Jeongbin Park, Kayeong Lim, Jin-Soo Kim, and Sangsu Bae. Cas-Analyzer: an online tool for assessing genome editing results using NGS data. *Bioinformatics* 33, 286-288 (2017).
- Jeongbin Park and Sangsu Bae. Cpf1-Database: web-based genome-wide guide RNA library design for gene knockout screens using CRISPR-Cpf1. *Bioinformatics* 34, 1077-1079 (2018).

1.4.2 Work solely done in Heidelberg

- Jeongbin Park, Liam Childs, Daesik Kim, Gue-Ho Hwang, Sunghyun Kim, Sang-Tae Kim, Jin-Soo Kim, and Sangsu Bae. Digenome-seq web tool for profiling CRISPR specificity. *Nature methods* 14 (6), 548 (2017).
- Sang-Tae Kim*, Jeongbin Park*, Daesik Kim, Kyoungmi Kim, Sangsu Bae, Matthias Schlesner, and Jin-Soo Kim, Response to “Unexpected mutations after CRISPR–Cas9 editing in vivo”, *Nature Methods* 15, 239–240 (2018).

- Gue-Ho Hwang*, Jeongbin Park*, Kayeong Lim, Sunghyun Kim, Jihyeon Yu, Eunchong Yu, Sang-Tae Kim, Roland Eils, Jin-Soo Kim, and Sangsu Bae. Web-based design and analysis tools for CRISPR base editing. *BMC Bioinformatics* 19, 542 (2018).

*: Shared first authors.

Chapter 2

Omics data analysis pipeline development

2.1 Introduction

2.1.1 Big omics data analysis using bioinformatics pipelines

Recent technological advances of omics techniques in the past decade, especially next-generation sequencing [63] or single-cell omics methods [64], allowed researchers to generate big data in a cost-efficient manner. As a result, the bioinformatics data size is getting larger and larger, and the scale of such data is easily going “beyond petabytes (PB) or even exabytes (EB)” [65]. Although there are some general-purpose programs, e.g. spreadsheet programs or statistical analysis packages, that can be used to process some of such data, it is challenging to address sophisticated problems without dedicated computational tools.

The computational tools are usually developed to be highly optimized to do a specific task [66]. For example, there are tools dedicated to aligning short DNA reads from the sequencing machine to the reference genome (sequence aligners) [67], or calling the variants of an individual sample against the reference genome (variant callers) [68]. Here, by simply connecting the output and input of the two different tools - the sequence aligner and the variant caller - one can make a workflow to call variants from the raw data from the sequencing machine. By doing that, one does not have to run the two tools separately, but the final result can be easily produced by just running the workflow with the raw data. Also by integrating some parameters which do not need to be changed so often as default parameters within the workflow, it is possible to avoid setting the complex set of parameters for different tools every time, which can reduce human error when analyzing many samples. Overall, making such workflows can reduce the time and effort spent on repetitive bioinformatics analyses, and, as many intermediate steps are handled programmatically and are thus less prone to human error, making them more reproducible. Such workflows are called bioinformatics pipelines.

In the real world, things are usually more complex. For the example of the simple variant

calling pipeline described in the above paragraph, there are several different aligners and variant callers, claiming that they are the best among the available tools out there. It is also important to select the tools that meet one's requirements, for example, certain output formats or hardware requirements. Therefore, it is needed to do benchmark the available tools and select the ones most suitable for one's situation. Moreover, the selection of the tools is not the only problem – there is also a variety of the reference data available. For example, there are several different versions of reference sequences out there, and it is needed to select the one based on the type of organisms. There also might be some exceptional cases that one should consider, e.g. sample contamination, swap, bad quality, etc. Therefore, the pipeline should have a quality control step to identify such cases. In the end, the pipeline has to be more complex to be flexible enough to account for many situations and needs.

In addition, it is as much as important to optimize the pipelines to harnesses the full capacity of accessible hardware, to perform analysis in a reasonable time frame. Thanks to the recent advances in technologies, the cost to produce bioinformatics data is getting lower, and as a result, the data size is getting bigger, leaving challenges in terms of optimization. Currently, many research institutes nowadays have a dedicated cluster system capable to run a lot of analyses in parallel [69], also big companies like Amazon or Google even offer cloud-based cluster systems, which can be used for bioinformatics analysis in a relatively cheap prices [70, 71]. Despite the availability of such systems, each system has its own set of parameters and requirements, therefore there is no bioinformatics tool that can work with every type of such system. Instead, the bioinformatics pipeline can be configured to properly use such a system, for example submitting certain tasks in parallel in different cluster nodes, to finish analysis in a short time. In short, having good flexible pipelines comprised of the right tools for the research is crucial for successful bioinformatics analyses [72].

Although it is important to have workflows for the specific system requirements of each institute, it would be so hard to collaborate between institutes if the input/output file format of such workflows is completely different from each other. Therefore, several standard file formats have been suggested for the bioinformatics data, especially for the next-generation sequencing data [57, 73, 74], frequently being used for pipeline development. Not only for the data but also there have been several efforts to achieve portability of the pipeline itself, by developing portable workflow management system [75–78] or even defining a standard language to write pipelines [75, 78, 79]. Based on them, many pipelines have been available in public [80, 81].

In this chapter, I describe the development of two new bioinformatics pipelines based on the existing pipelines at the DKFZ: 1) small somatic-like variant calling pipelines without matched control (in short, no-control small variant calling pipelines), and 2) single-cell RNA sequencing data preprocessing pipeline. Thanks to the workflow management system Roddy (<https://rodody-documentation.readthedocs.io/>), both pipelines are optimized to quickly process big sequencing

data (e.g. whole genome sequencing data in a petabytes scale cohorts, or single-cell transcriptome data from millions of single cells) by using the dedicated cluster system at the DKFZ, to finish analysis in several days.

2.1.2 Small variant calling without matched controls

Variant calling refers to a process in which genetic variants are identified based on sequencing data. Here, “variants” refers to the difference between the sequencing data to a reference genome, where the reference genome can be either the “standard” genome of an organism based on the consensus of the sequencing data, or a genome assembled from sequencing data obtained from a different, non-malignant sample from the same individual. Among them, the differences between the standard genome and the non-malignant sample are called germline variants, whereas the difference between the non-malignant sample to the malignant sample is called somatic variants. The type of genetic variants can be largely classified by 3 different categories based on the size of the variant; 1) single nucleotide variants (SNVs) (the size is 1 nucleotide) or multiple nucleotide variants (MNVs) (two or more SNVs in succession), 2) small insertions and deletions (indels) (the size is several nucleotides), 3) structural variants (SVs) or copy number variants (CNVs) (the size is more than several nucleotides). Here I define both SNVs and indels as “small” variants since they are as small as several nucleotides and can easily be called by existing variant calling software.

When it comes to cancer research, it is important to identify genetic variants in cancer tissue compared to the non-malignant tissue of the same individual (i.e. matched control), which are supposed to be the set of mutations where some of them drive carcinogenesis. Therefore, somatic variant calling pipelines are incorporated to call the cancer-specific somatic variants. However, not every cancer sample always has matched control. For example, the tumor/control sample was swapped, or the tumor was sequenced twice instead of sequencing control. Moreover, although the price of sequencing is getting lower and lower by the time, still whole-genome sequencing is expensive for the ones who have a limited budget – if one does not have to sequence matched controls, it would be possible to include twice as many cancer patients with the same budget.

Here I describe a “no control” small variant calling pipelines that simulate somatic SNV and indel calling without requiring a matched control, by subtracting common variants (in some publicly available common variants databases) from a germline variant calling result. The pipelines are developed as an extension of the existing in-house somatic variant calling pipeline at the DKFZ, so that it can be easily integrated with our in house workflow management systems, Roddy (<https://roddy-documentation.readthedocs.io/en/latest/>) and the One Touch Pipeline (OTP) [82].

2.1.3 Single cell RNA sequencing data analysis

Thanks to the recent advance of microfluidics techniques, several different methods have been developed which are capable of isolating single cells and analyze individual cells [83,84]. With the development of the techniques, the single-cell transcriptomics shed light to identify the heterogeneity of the individual cell types in tissue, which could not be revealed by the conventional bulk RNA sequencing methods [85].

Despite the difference of such single-cell isolation techniques, these methods commonly use small molecular barcode in the first read of paired-end sequencing (R1) to identify the individual single cell. To analyze such single-cell RNA sequencing (scRNA-seq) data, it is firstly needed to identify single cells using this barcode, and do quantification. Although there have been several attempts to develop a pipeline to do downstream analysis based on the “count matrix” (i.e. the matrix of the quantified mRNAs per single cell) [86,87], but so far still the limited number of pipelines have been developed to generate such count matrix from the raw sequencing data. Many single-cell isolation techniques are being developed by commercial companies and sometimes accompanied by a dedicated pipeline for this [88], but such pipelines are not designed for analysis of sequencing data produced by different machines. Thus, one has to learn how to use various software when they want to use a different technique.

Therefore, I developed a highly flexible single-cell RNA sequencing data analysis pipeline, that produces a gene expression matrix from single-cell RNA sequencing data with any library design. The pipeline was initially developed to easily analyze data obtained by Fluidigm C1 and Wafergen machines but later extended to process single-cell RNA sequencing data from other machines. Since data processing of single-cell RNA sequencing data is quite similar to that of bulk RNA sequencing data, the pipeline was developed as an extension of the existing DKFZ RNA sequencing pipeline, so that the pipeline can be harmonized with our in house workflow management systems, Roddy (<https://roddy-documentation.readthedocs.io/en/latest/>) and the One Touch Pipeline (OTP) [82].

2.2 Results

2.2.1 No control variant calling pipeline

Both no control SNV and indel pipelines accept an aligned sequencing data as a BAM (binary sequence map) file as an input. After the initial germline variant calling, both pipelines use three public databases, dbSNP [89], ExAC [90], and EVS (<http://evs.gs.washington.edu/EVS/>), and an in-house control dataset (containing 280 controls), to subtract common variants from the identified germline variants. Finally, the remaining variants are reported as a variant calling format (VCF) file.

By default, both pipelines subtracts the common variants defined in dbSNP with ‘COMMON=1’ tag, which refers to the MAF (minor allele frequency, the fraction of population who have the variants among the whole population) value is more than 0.01 in at least one of the 5 major populations in 1000 genome project as following: African, Ad Mixed American, East Asian, European, and South Asian. Additionally, the variants found in the ExAC database with more than 0.1 %, in the EVS database with more than 1 %, and in the in-house control dataset with more than 2 % were removed. Among the removed variants, the ones in the OMIM [91] record are not removed (“rescued”) since they could be clinically important although they are common in the population. Such criteria can be adjusted as parameters of the pipeline.

This strategy was tested with 4 malignant lymphoma cancer samples from ICGC-MMML-Seq project [92] (<https://icgc.org/icgc/cgp/64/345/53049>). After the subtraction of the common variants, the SNVs found in coding regions (functional SNVs) were 200 - 250, which is comparable to 100 - 150 SNVs found with a matched control. Also, the number of functional SNVs that could not be identified by no-control workflow was 16, 11, 17, and 48, respectively. The number of functional indels was 20 - 30, which is also comparable to 5 - 10 indels found with control. The number of functional indels could not be identified by no-control workflow was 4, 1, 2, and 1, respectively.

2.2.2 Use case: No control variant calling for the ovarian cancer samples

The development of no control pipelines was started to call somatic-like variants from 10 different ovarian cancer samples without matched controls, which were cultured in two different environments: in 2D and 3D. Dr. Julia Jabs, who performed the wet-lab experiments in this project, found that the drug response was different in the two cell-culturing environments by *DeathPro*, an automated microscopy-based assay which can resolve cell death and proliferation inhibition in both 2D and 3D [93]. One of the main questions of the project was to know which environment would reflect the expected phenotype by the mutations profile discovered by whole-genome sequencing analysis. Our analysis revealed that the TP53 gene, which is one of the most commonly mutated genes in many cancers, was mutated in 9 of 10 samples that we have, but overall few SNVs or indels were detected in other genes. Instead, we found that a numerous copy number variations (CNVs) were detected in all cancer samples using other no control pipelines, Sophia [95] and ACEseq [96] (Figure 2.1). Based on the result, we found that the homologous recombination deficiency (HRD) score (i.e. the number of loss of heterozygosity (LOH) regions >15 Mbp) [97] showed high correlation with Paclitaxel response in the cultured organoids.

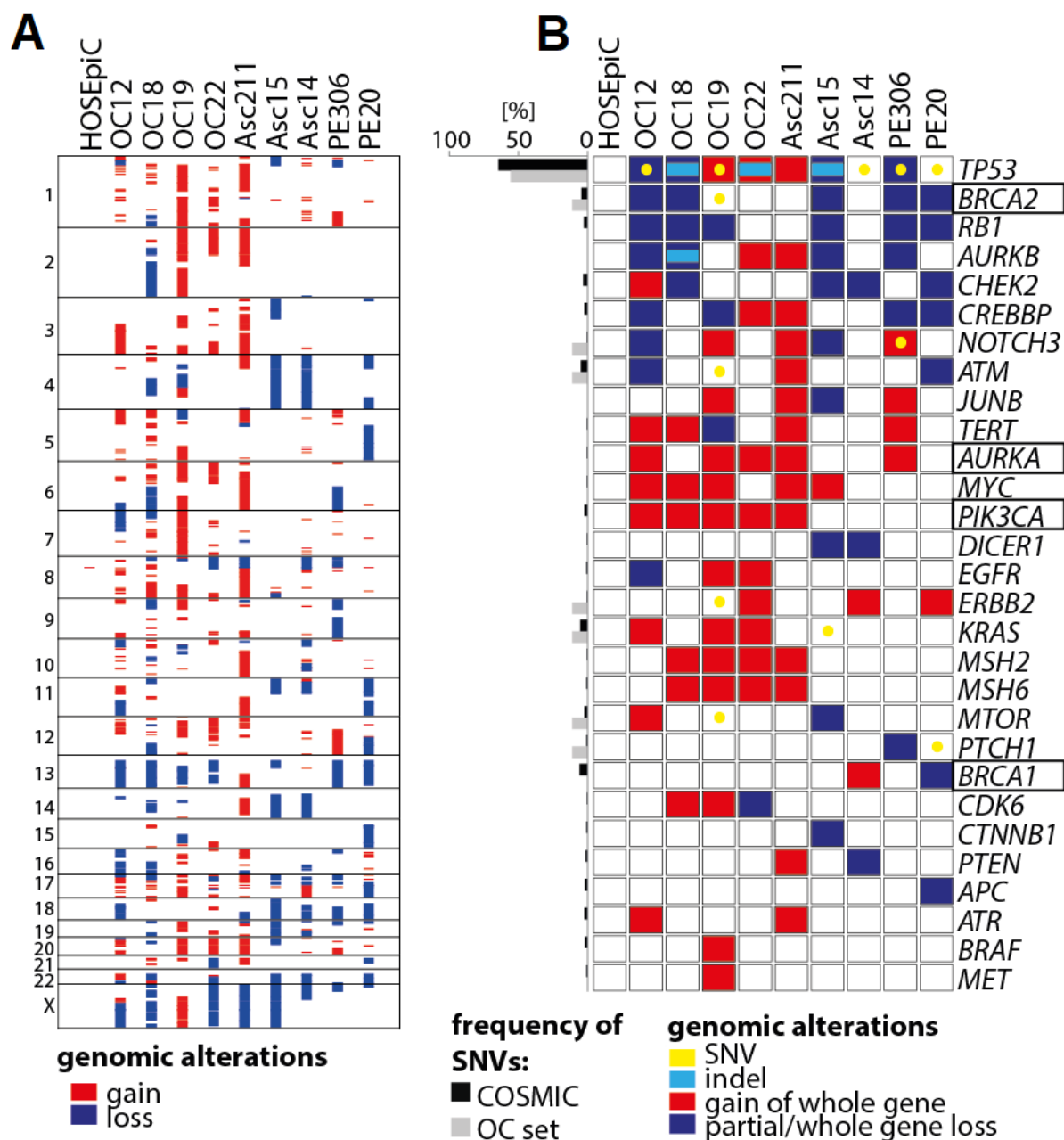


Figure 2.1. Ovarian cancer (OC) samples having numerous copy number variations (CNVs) but few small somatic variations.

(Note: This figure was jointly produced with Dr. Julia Jabs, and later published as a part of Jabs et al. (2017), in Figure 5 [93].)

(A) OC samples contain lots of CNVs compared to a normal ovarian sample (HOSEpic); (B) The sets of frequently mutated genes (selected from COSMIC [94] and ICGC databases) in OC samples also majorly harbor CNVs, compared to the few small variations except *TP53*.

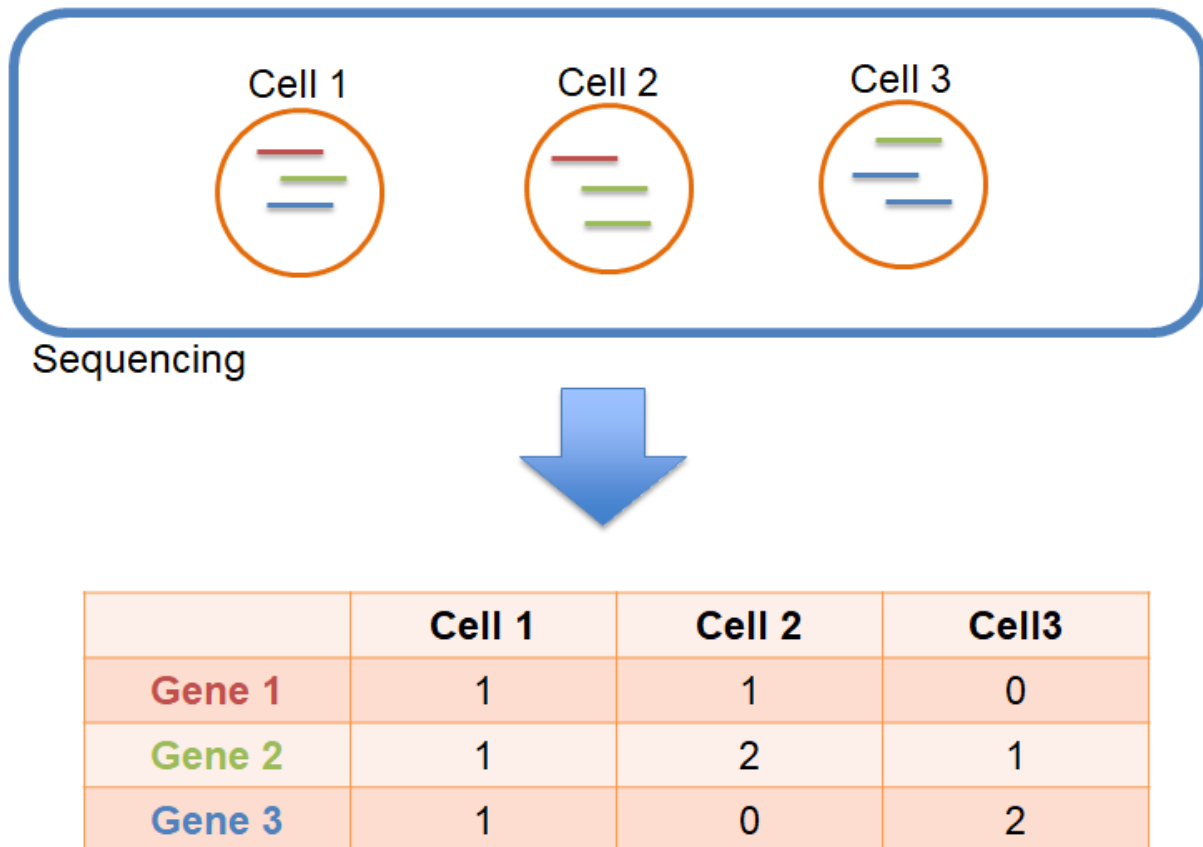


Figure 2.2. Overview of single-cell RNA sequencing data preprocessing workflow.

Input FASTQ file (blue box) contains all of the reads in one file. The reads are demultiplexed using the cell barcodes and then quantified to produce the count matrix. In the figure, each colored gene correspond to each colored mRNA in the cells.

2.2.3 Single cell RNA-seq data preprocessing pipeline

The development of the pipeline was begun to process raw data generated by the Wafergen machine, where the experiments were done by Dr. Stephan Tirier. The goal of the preprocessing workflow is to process Fastq files from the sequencing machine and then finally create a matrix which contains gene expression values of each cell (Figure 2.2). The workflow is designed to be fully scalable so that it can even process Fastq files containing millions of cells. The processing steps can be classified into 3 steps: demultiplexing, alignment/quantification, and quality control.

In demultiplexing step, cell barcode and UMI in Read 1 (Figure) are processed. In addition, poly N tails and remaining fragments of PCR adapter in Read 2 are trimmed, and after that Read 2 is filtered by quality. Finally, the trimmed reads are demultiplexed.

The alignment/quantification step was implemented by DKFZ HIPO team (led by Dr. Naveed Ishaque) for bulk RNA sequencing data. The read alignment is powered by STAR, and the parameters for STAR have been fine-tuned to provide accurate alignment results. Since the single-cell RNA sequencing workflow was developed as an extended pipeline of the bulk RNA sequencing workflow, it uses the same parameter set for the alignment. After that quantification is performed, and finally the expression matrix is generated. Afterward, a quality control script is executed and generates QC plots for users (Figure). Since we still want to adjust QC parameters manually for each sample, the tool only shows QC plots but does not actually perform any filtering based on it.

2.2.4 Use case: Pheno-seq project

With Dr. Stephan Tirier, I have worked on sequencing data analysis of dissected microtissue (miti), named as Pheno-seq [98]. The idea of Pheno-seq is to dissect microtissue cultured in 3D [99–101], by using the same microfluidics-based apparatus and the sequencing machine as for single-cell RNA sequencing (scRNA-seq). Since the morphological difference determined by imaging microtissues infers important information on the cell heterogeneity, we would like to combine this information together with the sequencing data. In addition, since each microtissue is comprised of multiple cells, we can capture more reads in Pheno-seq than scRNA-seq, and thus Pheno-seq provides an even more robust result than scRNA-seq.

In this project, we tested Pheno-seq with a breast cancer cell line (MCF10CA) and then tried to apply this approach to patient-derived colon cancer samples. Additionally, we also performed scRNA-seq of the same samples and compared the results.

Firstly Pheno-seq was tested using MCF10CA cell lines. Here, the transition of the epithelial cells, which is known to occur in the initiation of metastasis [102], can be observed by imaging microtissue (Figure 2.5A). This phenotype difference was also reflected in the sequencing data

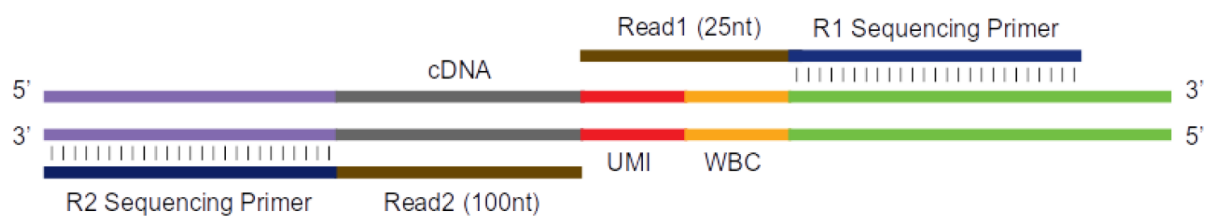


Figure 2.3. NextSeq sequencing library structure of Wafergen data.

NextSeq sequencing library structure of Wafergen data. Read1 contains unique molecular identifier (UMI) and well barcode (WBC), and Read2 contains cDNA.

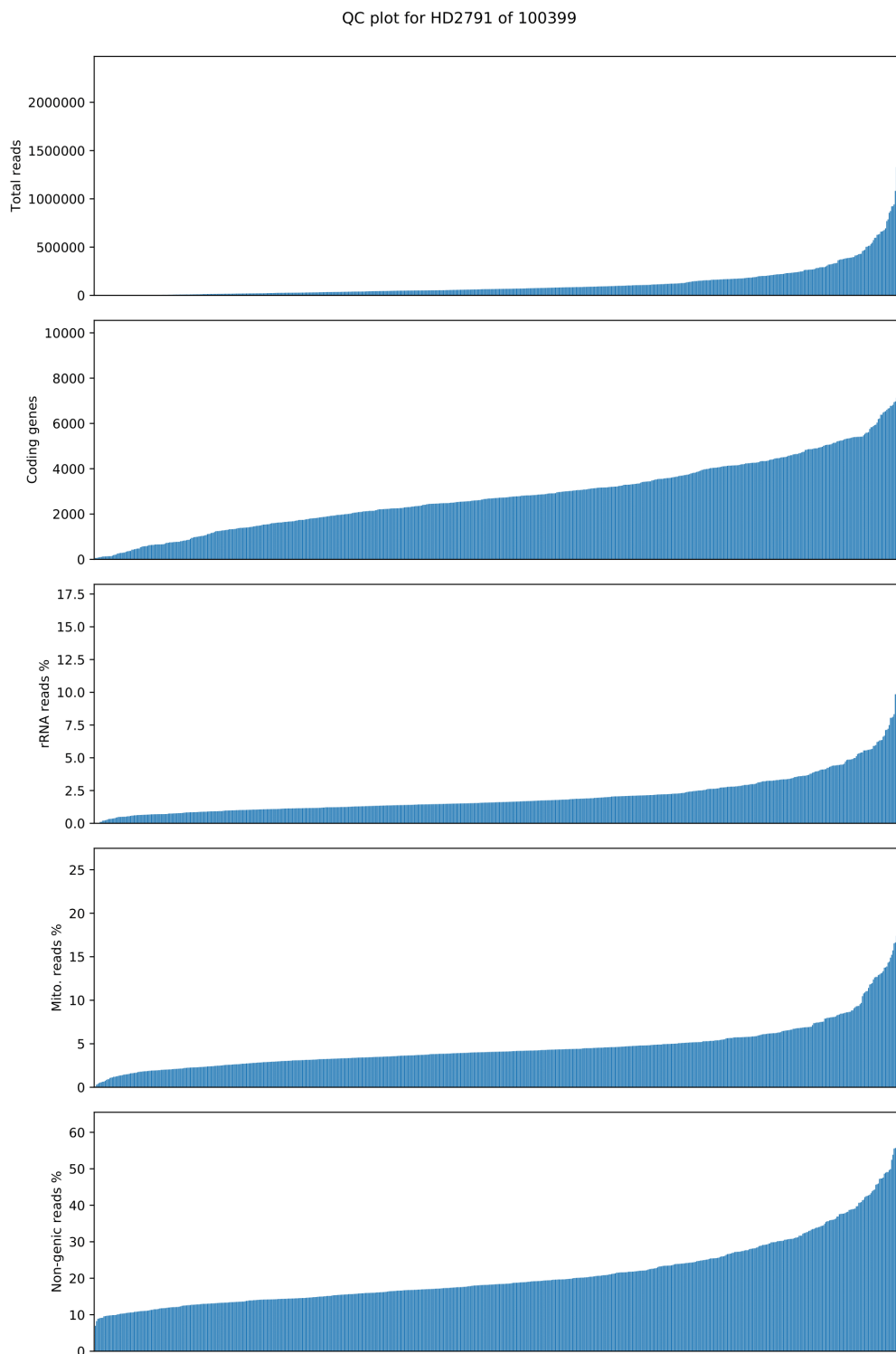


Figure 2.4. Quality control plots from one example sample provided by the preprocessing workflow.

Plot for total reads, number of coding genes (expression > 0), reads aligned to rRNA, reads aligned to mitochondria, and reads cannot be mapped to a gene (from top to bottom).

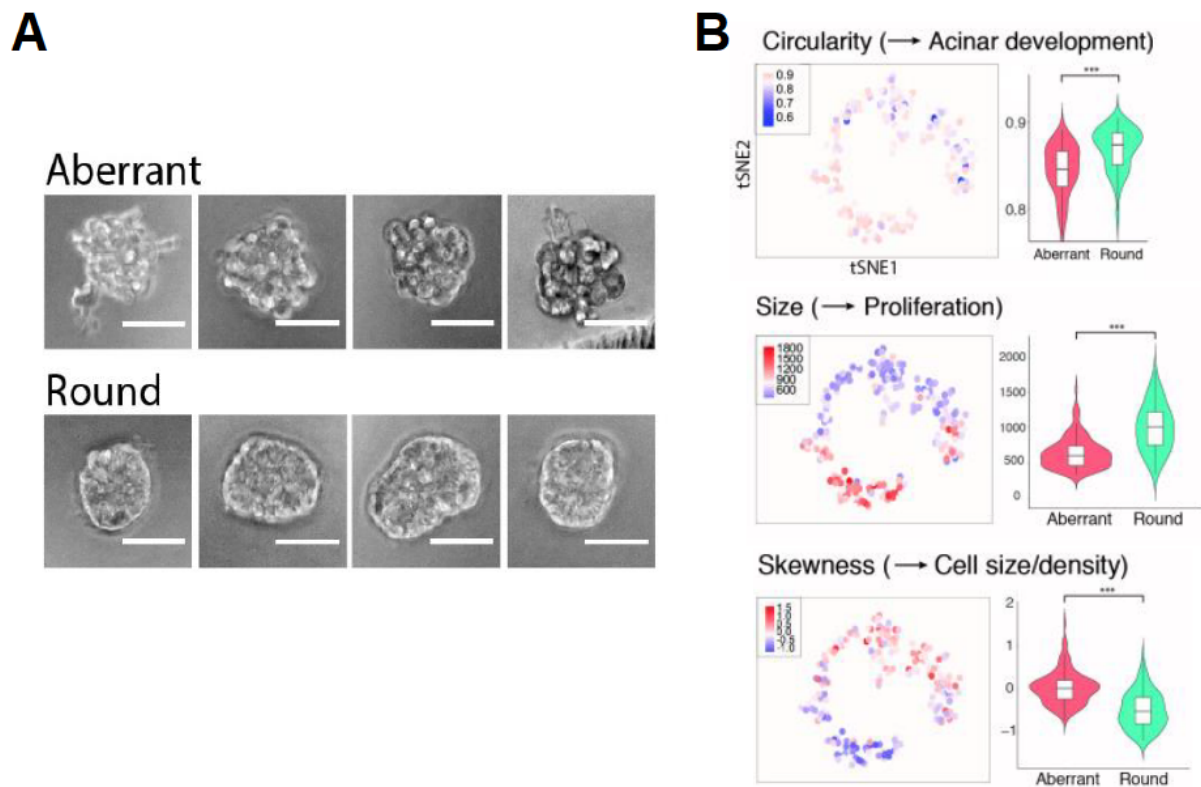


Figure 2.5. Pheno-seq analysis of MCF10CA cell line.

(Note: Panel A was drawn by Dr. Stephan Tirier, and panel B was jointly produced with Dr. Stephan Tirier. This figure was presented as a part of the poster presentation at the winter PhD poster presentation at the DKFZ by Dr. Stephan Tirier.)

(A) Brightfield images of aberrant (invasive) and round microtissues of breast cancer cell line MCF10CA. Scale bar is 50 μm ; (B) 2D t-SNE embedding of the microtissues, colored by each imaging features (circularity, size, and skewness). The microtissues were clustered into two clusters with k-means clustering, and the distribution of imaging features within each cluster was shown as violin plots.

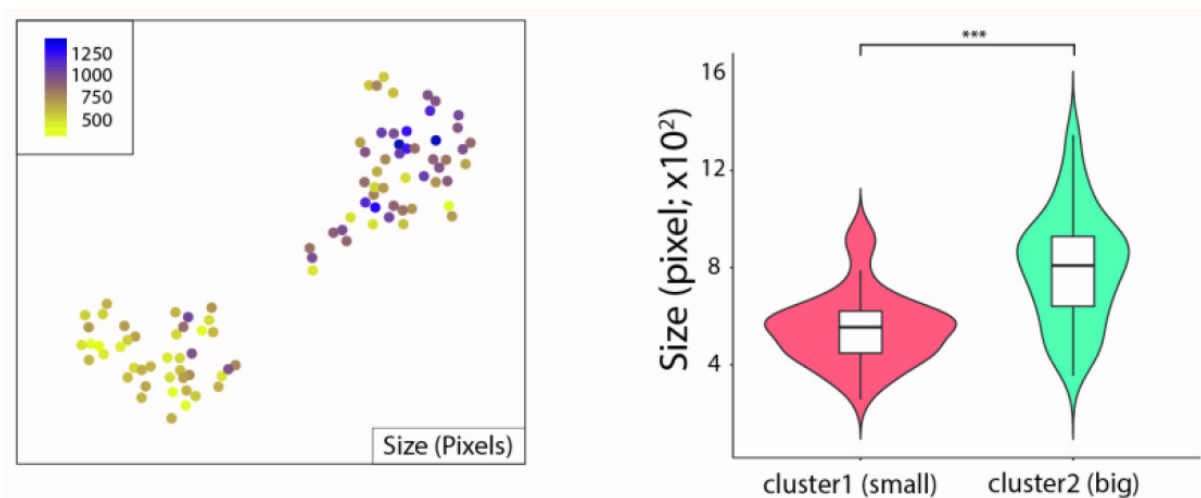


Figure 2.6. Pheno-seq result of colorectal sample.

(Note: This figure was produced jointly with Dr. Stephan Tirier. This figure was presented as a part of poster presentation at the winter PhD poster presentation at the DKFZ by Dr. Stephan Tirier.)

(Left) 2D t-SNE embedding of the microtissues from colorectal sample; (Right) Distribution of the size of the microtissues within the 2 clusters identified by k-means clustering.

and appeared, as two distinct clusters in the t-SNE plot. We also compared the two clusters with several imaging features that can be used to discriminate aberrant and round cells: circularity related to acinar development, size related to proliferation, and skewness related to cell size/density of the microtissues. The image features were nicely matched into the two clusters the t-SNE plot, which shows a good agreement with the clusters determined by Pheno-seq (Figure 2.5B). The Pheno-seq system was also applied to a colon cancer patient sample. From the sample, tumor-initiating cell (TIC), transient amplifying cell (TAC), and postmitotic cell were dissociated and cultured individually to form two different types of microtissues and then sequenced. The data showed two distinctive clusters and the image feature (size) matches well with the clusters that we found, which makes a good agreement with expectation (Figure 2.6).

2.3 Discussion

Overall, the no control somatic small variant calling pipelines are tremendously helpful when the matched germline control is not sequenced or not available due to several reasons. After the subtraction of millions of common variants from public variant databases, most of the germline variants were successfully filtered, resulting in the number of private, rare and somatic-like variants to be reasonably small. It was validated that the no control pipelines can identify most of the somatic variants found by somatic variant calling with a matched control. Since the remaining variants can still contain some germline variants, the no control workflow is especially useful for a cohort level settings, to find frequent driver mutations in a cohort. After successful implementation of the pipeline in the ovarian cancer cohort, it was also used for several other projects as following: Juvenile myelomonocytic leukemia (JMML) [103], Squamous-cell skin cancer (cSCC) [104], and Burkitt lymphoma projects [105].

The single-cell analysis pipeline was developed to quantify the mRNAs in every single cell. The pipeline as successfully used to analyze not only single-cell RNA sequencing data but also Pheno-seq data which is sequencing organoids instead of single cells. The pipeline was also used for the following two different projects: Glioblastoma project [106], and the human colorectal cancer project.

2.4 Contributed publications

- Julia Jabs, Franziska M Zickgraf, Jeongbin Park, Steve Wagner, Xiaoqi Jiang, Katharina Jechow, Kortine Kleinheinz, Umut H Toprak, Marc A Schneider, Michael Meister, Saskia Spaich, Marc Sütterlin, Matthias Schlesner, Andreas Trumpp, Martin Sprick, Roland Eils, and Christian Conrad, Screening drug effects in patient-derived cancer cells links organoid responses to genome alterations. *Molecular Systems Biology* 13 (11), 955

(2017).

- Rabea Wagener, Cristina López, Kortine Kleinheinz, Julia Bausinger, Sietse M. Aukema, Inga Nagel, Umut H. Toprak, Julian Seufert, Janine Altmüller, Holger Thiele, Christof Schneider, Julia Kolarova, Jeongbin Park, Daniel Hübschmann, Eva M. Murga Penas, Hans G. Drexler, Andishe Attarbaschi, Randi Hovland, Eigil Kjeldsen, Michael Kneba, Udo Kontny, Laurence de Leval, Peter Nürnberg, Ilske Oeschies, David Oscier, Brigitte Schlegelberger, Stephan Stilgenbauer, Wilhelm Wössmann, Matthias Schlesner, Birgit Burkhardt, Wolfram Klapper, Elaine S. Jaffe, Ralf Küppers, and Reiner Siebert. IG-MYC+ neoplasms with precursor B-cell phenotype are molecularly distinct from Burkitt lymphomas. *Blood* 132 (21): 2280-2285 (2018).
- Manuel Rodríguez-Paredes, Felix Bormann, Günter Raddatz, Julian Gutekunst, Carlota Lucena-Porcel, Florian Köhler, Elisabeth Wurzer, Katrin Schmidt, Stefan Gallinat, Horst Wenck, Joachim Röwert-Huber, Evgeniya Denisova, Lars Feuerbach, Jeongbin Park, Benedikt Brors, Esther Herpel, Ingo Nindl, Thomas G. Hofmann, Marc Winnefeld, and Frank Lyko. Methylation profiling identifies two subclasses of squamous cell carcinoma related to distinct cells of origin. *Nature Communications* 9, 577 (2018).
- Daniel B. Lipka, Tania Witte, Reka Toth, Jing Yang, Manuel Wiesenfarth, Peter Nöllke, Alexandra Fischer, David Brocks, Zuguang Gu, Jeongbin Park, Brigitte Strahm, Marcin Wlodarski, Ayami Yoshimi, Rainer Claus, Michael Lübbert, Hauke Busch, Melanie Boerries, Mark Hartmann, Maximilian Schönung, Umut Kilik, Jens Langstein, Justyna A. Wierzbinska, Caroline Pabst, Swati Garg, Albert Catalá, Barbara De Moerloose, Michael Dworzak, Henrik Hasle, Franco Locatelli, Riccardo Masetti, Markus Schmugge, Owen Smith, Jan Stary, Marek Ussowicz, Marry M. van den Heuvel-Eibrink, Yassen Assenov, Matthias Schlesner, Charlotte Niemeyer, Christian Flotho, and Christoph Plass. RAS-pathway mutation patterns define epigenetic subclasses in juvenile myelomonocytic leukemia. *Nature Communications* 8, 2126 (2017).
- Stephan M. Tirier, Jeongbin Park, Friedrich Preußner, Lisa Amrhein, Zuguang Gu, Simon Steiger, Jan-Philipp Mallm, Teresa Krieger, Marcel Waschow, Björn Eismann, Marta Gut, Ivo G. Gut, Karsten Rippe, Matthias Schlesner, Fabian Theis, Christiane Fuchs, Claudia R. Ball, Hanno Glimm, Roland Eils, and Christian Conrad. Pheno-seq — linking visual features and gene expression in 3D cell culture systems. *Scientific Reports* 9, 12367 (2019).
- Teresa G Krieger, Stephan M Tirier, Jeongbin Park, Tanja Eisemann, Heike Peterziel, Peter Angel, Roland Eils, and Christian Conrad. Modeling glioblastoma invasion using human brain organoids and single-cell transcriptomics. Under review.

– Preprint available at BioRxiv, <https://doi.org/10.1101/630202>

- Martina K. Zowada, Stephan M. Tirier, Sebastian M. Dieter, Teresa Krieger, Ava Oberlack, Robert L. Chua, Mario Huerta, Foo W. Ten, Karin Laaber, Jeongbin Park, Katharina Jechow, Torsten Müller, Mathias Kalxdorf, Mark Kriegsmann, Katharina Kriegsmann, Friederike Herbst, Jeroen Krijgsveld, Martin Schneider, Roland Eils, Hanno Glimm, Christian Conrad, and Claudia R. Ball. Functional states in tumor-initiating cell differentiation in human colorectal cancer. Under review.

Chapter 3

Segmentation-free inference of cell types from *in situ* transcriptomics data

3.1 Introduction

Thanks to the recent advances of the single-cell RNA sequencing techniques [107], the profound heterogeneity of different types of single cells in tissues has been successfully revealed and has led to the birth of international consortia such as Human Cell Atlas (HCA) [108]. Linking such single-cell heterogeneity found by single-cell sequencing experiments in the spatial context at tissue level enabled to unravel the transcriptional heterogeneity of invasive cancer tissues [109], or the localization of different neuronal subtypes in the brain cortex or hypothalamus [110, 111].

The recent effort to measure the gene expressions in spatial context has been established via multiplexed fluorescence *in situ* hybridization (FISH) [110, 112, 113] or *in situ*/intact tissue sequencing [114–119] methods, which enabled simultaneous measurement of the localization of different mRNAs in a spatial context, which given rise to the international consortia like the SpaceTx consortium [120].

The current methods to find cell types in spatial context rely on cell segmentation algorithms, by quantifying the mRNAs in each cell segment and use this as the gene expression of the cells, and process the resulting count matrix as like processing single-cell sequencing data. The cell segmentation algorithms currently used usually rely on additional materials, such as landmarks of stained nuclei [121], cell membrane [122–124], or total mRNAs [110, 111]. Even with such additional materials, the segmentation algorithms are still limited by various imaging problems, due to the unclear cell borders, overlapping cells, signal intensity variation, and tiling artifacts [125], which leads to loss of cells in the end. The current segmentation-based cell-type calling algorithms also inherit such problems, which can affect the cell type calling efficiency.

In this chapter, I present a novel computational framework named Spot-based Spatial cell-type Analysis by Multidimensional mRNA density estimation (SSAM), which does not rely on

segmentation algorithms but only requires the location of mRNAs as an input. The accuracy of the segmentation-free cell type calling result, and the improvements of the resulting cell type map is discussed using two publicly available datasets, the osmFISH [110] and MERFISH [111] datasets. In addition, I show that the identification of a new cell type that has never been reported using the multiplexed smFISH dataset.

3.2 Results

3.2.1 The SSAM computational framework

The SSAM framework is comprised of 4 following steps: 1) Estimation of gene expression in space by Kernel Density Estimation (KDE), selection of representative locations, and normalization, 2) clustering analysis to find cell types, and 3) cell type map generation and 4) spatial domain analysis.

In the first step, SSAM estimates the gene expression at every lattice point on a periodic square lattice using Kernel Density Estimation (KDE) [126, 127] (Figure 3.1A). Here the 2D or 3D Gaussian kernel was used for the KDE, for mRNAs in 2D or 3D space, respectively. Since the spacing between the lattice points is the same for all axes, the lattice is analogous to an image – and each lattice point can be interpreted as a pixel (2D) or voxel (3D) in the image. After the estimation, SSAM stacks the estimated gene expression images, so that it becomes a stacked image with every pixel or voxel contains a signature of every gene instead of a single expression value. Since every pixel contains gene expression signature, which is a 1-dimensional vector, therefore the image is a field with gene expression vectors – therefore this can be called a vector field. Since the number of vectors in the vector field is often more than several million (for example, osmFISH dataset contains more than 7 million vectors in the vector field), SSAM selects ‘representative vectors’ that is a downsampled set of vectors in the vector field. For a downsampling strategy of vectors, SSAM focuses on the nature of estimated total gene expression signals after KDE. If the total gene expression is large in a pixel, this is more likely to be a cell, rather than the area between the cells. Moreover, the nature of KDE propagates the information to the nearby area, and the local maxima of the total gene expression within the neighboring pixels would be the locations of the pixels highly likely to be inside of cells. Since it most makes sense to select vectors originated from cells (Figure 3.2C), therefore, SSAM selects the local maxima of neighboring gene expressions as a default strategy to downsample vectors. The downsampled vectors are then further thresholded with the total gene expression to remove artifacts, and additionally can be restricted with an optional “input mask” within the region of interest (see Method detail section for further details). After that, the downsampled vectors and the whole vector field are normalized (also see Method detail

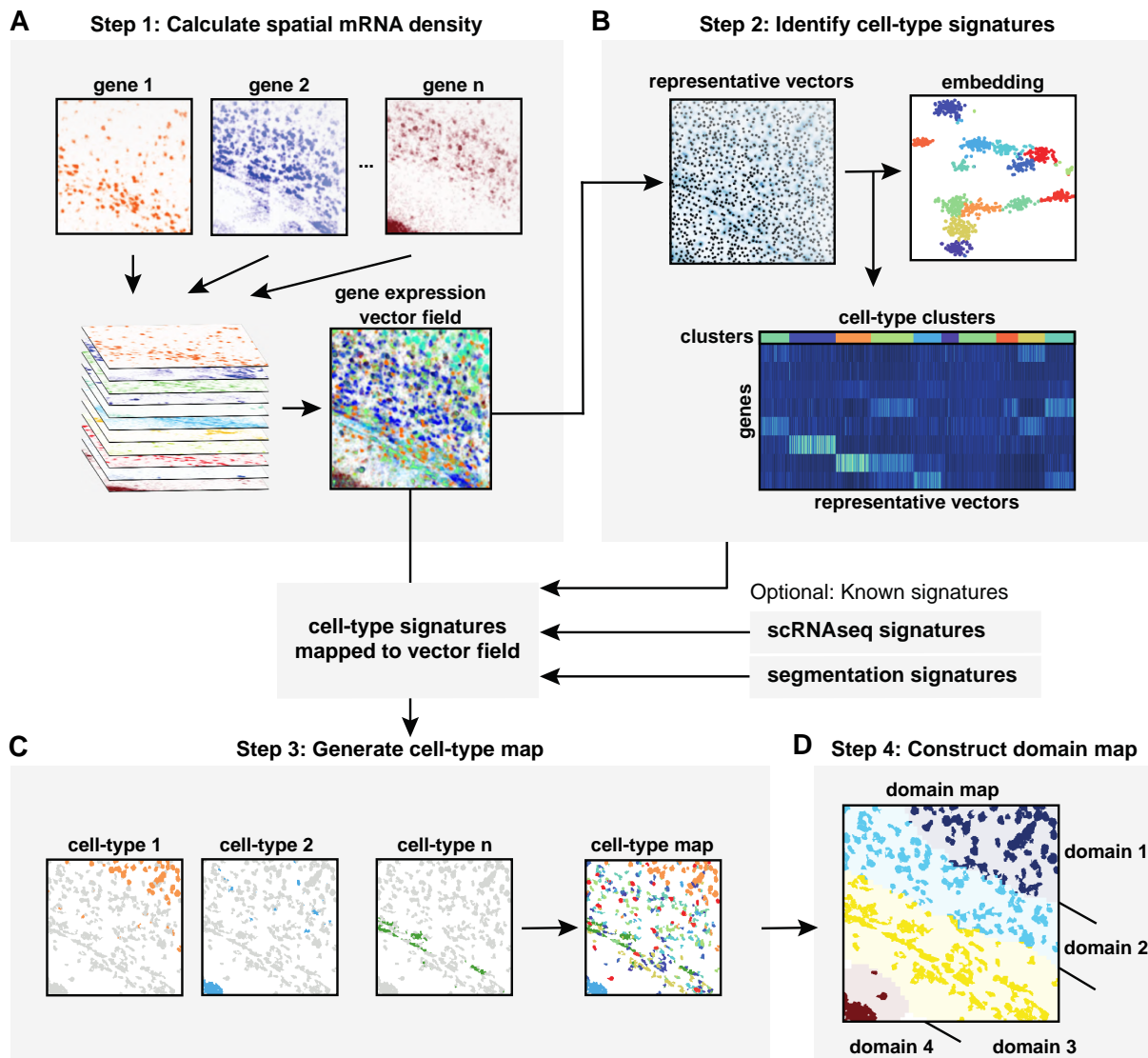


Figure 3.1. Schematic diagram of the SSAM computational workflow for cell type and tissue domain definition based on gene expression data.

(A) In step 1, the expression of each gene is calculated by estimating a spatial mRNA density by KDE. The calculated gene expressions are then stacked to form a gene expression vector field; (B) In step 2, the vectors in the vector field is downsampled to a fewer number of vectors to achieve computational feasibility. By default, the local maxima of the total gene expression are selected as the representative vectors. The selected vectors are then clustered, and the centroid of each cluster is considered to be the cell-type signature; (C) In step 3, the cell-type map is generated by calculating correlation between the cell-type signatures to the vector field, and then merging them by assigning the index of the cell-type with the maximum correlation to each pixel (see also Figure 3.3A); (D) In step 4, the domain map is constructed by sweeping a sliding circular/sphere window on the cell type map, and clustering the composition of the cell types in each window (see also Figure 3.4B).

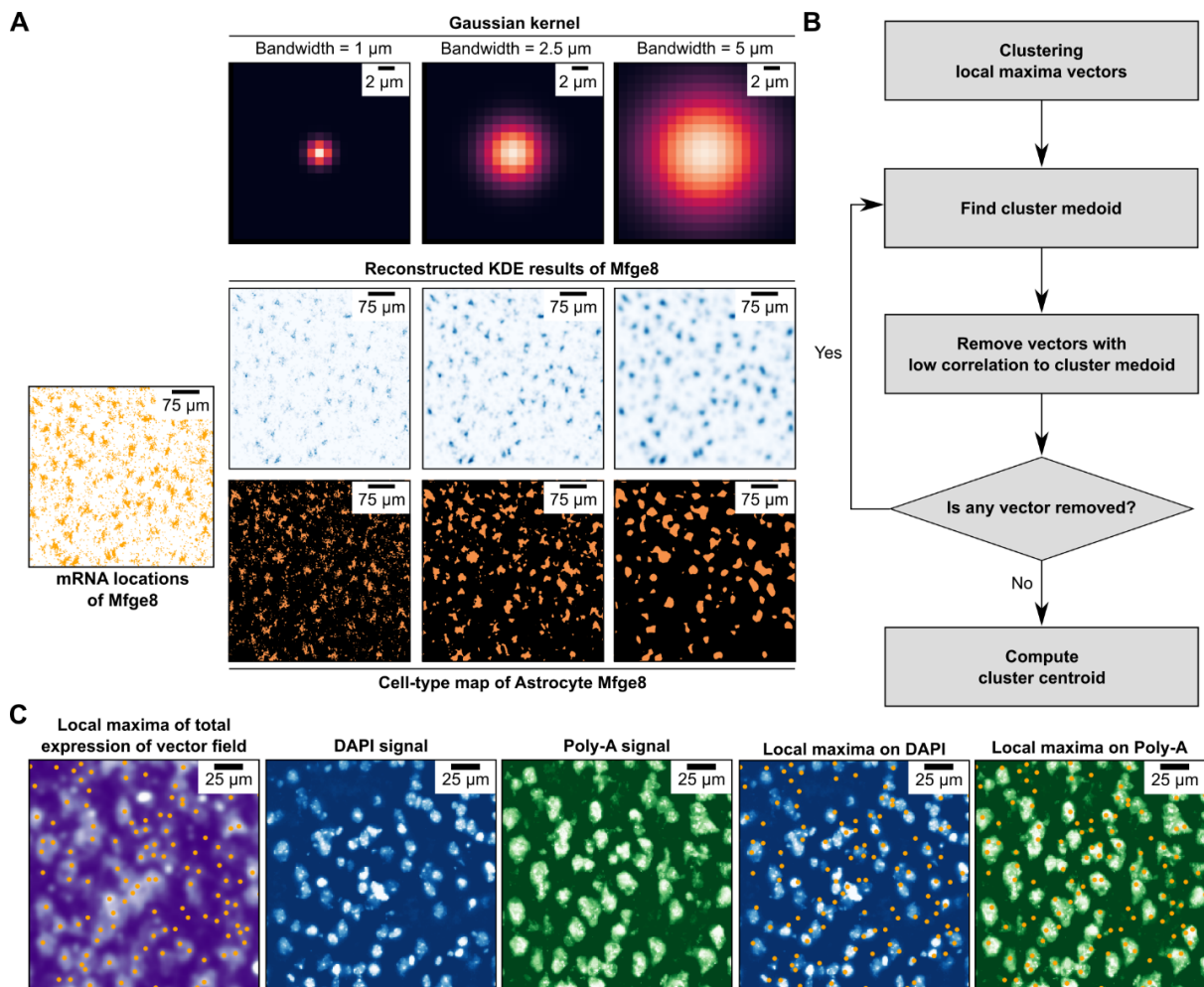


Figure 3.2. Identification of intracellular regions via mRNA density estimation.

(A) The effect of different bandwidth of Gaussian kernel on the estimation of gene expression, and the final cell-type map. Smaller bandwidth (1 μm) fails to fully reconstruct continuous gene expressions within the cell-like area, results loss of details in the cell-type map. Whereas larger bandwidth (5 μm) makes the gene expressions too much smoothed, which increases the chance to produce false gene expressions outside genes, resulting in larger blobs in the cell-type map. Still, the cell-type map shows not so much big difference in terms of the bandwidth difference in this range (1 μm and 5 μm), but extremely low or high bandwidths can produce an unreasonable result; (B) The strategy to exclude lowly correlated vectors from a cluster. Since Louvain clustering only divides the whole population but not excludes bad ones, it is needed to remove them manually; (C) The location of local maxima vectors highly overlaps with the regions with the Poly-A signal, proves the validity of local maxima selection strategy as a downsampling method for further downstream analyses.

section for further details). Since the whole vector field contains a lot of pixels outside of cells, therefore the parameters determined by the normalization of the downsampled vectors were used to normalize the vector field.

Next, SSAM can work in either guided mode or *de novo* mode, where the guided mode refers to the mode using predetermined cell-type signatures for further analysis, whereas the *de novo* mode uses a clustering algorithm to infer the cell types in the dataset. For guided mode, the following second step can be skipped.

In the second step (for *de novo* analysis), SSAM finds the cell types by clustering the downsampled vectors (Figure 3.1B). Here clusters are detected as the communities found by the Louvain community detection algorithm on the shared nearest neighbor (SNN) network based on the vectors, which is the same clustering method implemented in Seurat [86] (see Method detail section for further details). Although the downsampling in the second step removed the majority of vectors outside of cells, still it is possible that the local maxima can be selected outside of the cell area due to the noises. Therefore, after clustering the vectors that are not close to the cluster medoid are excluded from the clusters (Figure 3.2B). After that, the centroids of each cluster are calculated (i.e. the unweighted mean of the gene expression of the vectors in each cluster) and regarded as the representative signatures of each cell type. Here in case, some clusters which are mapped to artifacts in the image can be manually removed, and also the clusters have a very similar gene expression signature that can be manually merged based on the biological background knowledge. To make this easier, SSAM supports the creation of ‘diagnostic plots’ which shows useful information (see Method detail section for further details).

In the third step, each cluster centroid is mapped to spatial context by calculating Pearson’s correlation coefficient between the centroid to every pixel in the vector field (Figure 3.1C). Here each pixel is labeled as the label of the centroid with the highest correlation coefficient, in other words, the labeled images (i.e. cell-type map of each cell type) shows the spatial organization of each cell type, and these images are merged into a single image which is called cell-type map (Figure 3.3A). In the case of the generation of a cell-type map, it is possible to correlate foreign cell-type signatures, such as the ones found by the segmentation-based analysis or single-cell RNA sequencing signatures to the vector field to create a cell-type map. This procedure is called ‘guided mode’ SSAM, and the clustering analysis is called ‘*de novo* mode’ SSAM.

In the fourth step, SSAM shows the domain structure of the tissue using the cell type map, based on the cell type map found in the third step (Figure 3.1D). For this, a circular (or spherical) window with radius comparably larger than the average size of the cells is swept in the image, and then the proportion of the pixels of each cell type is clustered by agglomerative clustering to find the area that has the same proportion of cell types on tissue (Figure 3.3B).

SSAM is firstly demonstrated with two publicly available datasets, the somatosensory cortex (SSp) imaged by osmFISH and the hippocampal preoptic area (POA) imaged by MERFISH,

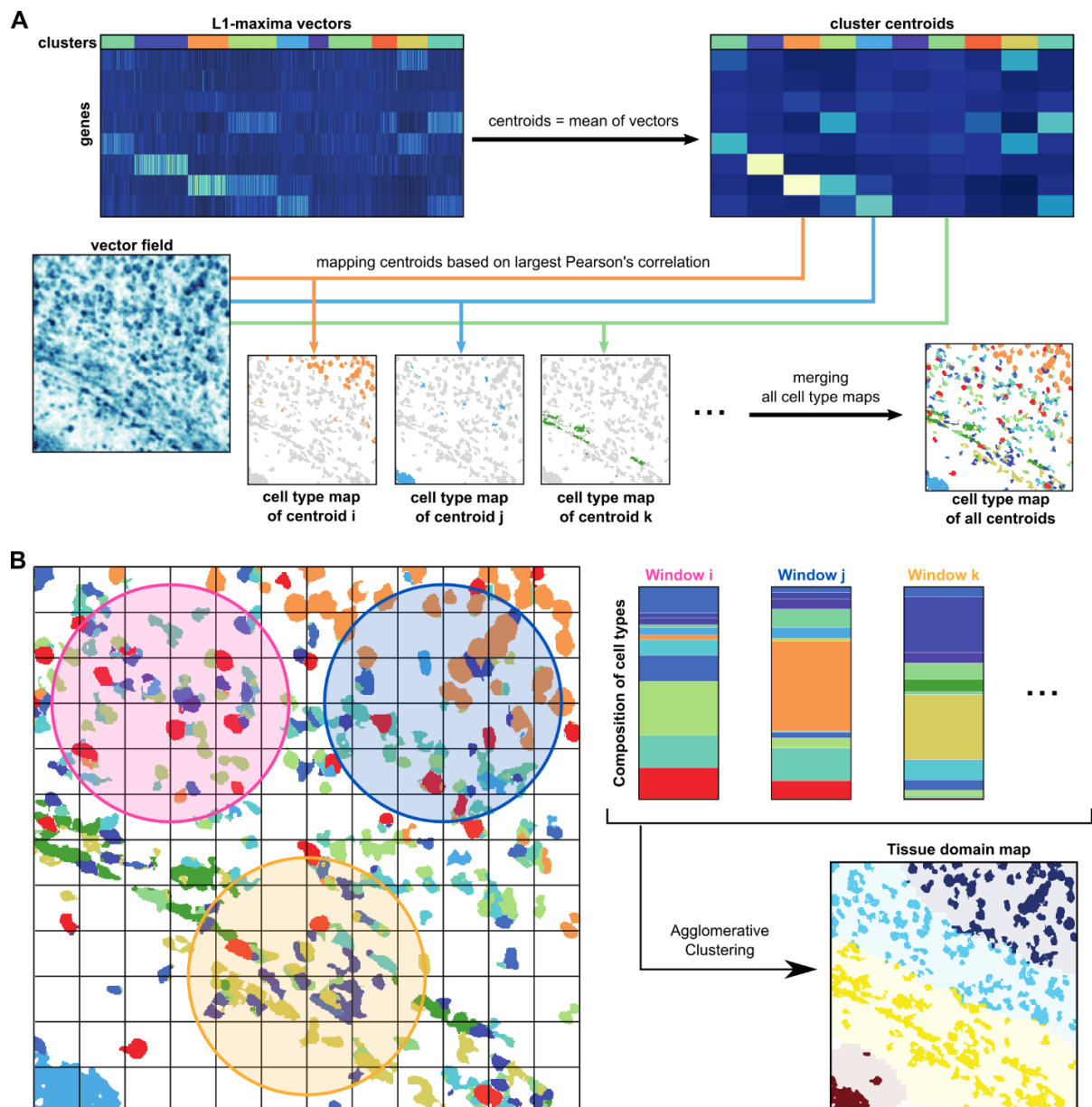


Figure 3.3. Generation of the cell-type map and the domain map.

(A) After clustering, the centroids (i.e., the unweighted mean of gene expressions of the vectors in the cluster) are calculated and regarded as the cell-type signatures. Each cell-type signature is then mapped to the vector field, by calculating Pearson's correlation (by default, a different correlation measure can also be used instead) between the cell-type signature to every vector in the vector field. For each pixel, the cell-type which has the maximum correlation is assigned. Finally, the final cell-type map is generated by merging the assigned pixels per cell-type; (B) On the cell-type map, a circular (or spherical) sliding window larger than the average cell size is swept, and the composition of cell types (as a composition of the number of pixels assigned to each cell type) of each window is clustered to generate a tissue domain map. In this chapter, Agglomerative clustering was used for all three datasets.

and then applied to newly generated adult mouse visual cortex (VISp) imaged by multiplexed smFISH.

3.2.2 Analysis of mouse somatosensory cortex (SSp) imaged by osmFISH

The robustness of SSAM was tested using the publicly available mouse brain somatosensory cortex dataset (SSp), imaged by osmFISH [110] (Figure 3.4, 3.5, 3.6, 3.7). The dataset contains 33 genes in 2D plane (x and y coordinates of mRNAs), each gene was sequentially measured individually at each image.

From the location of mRNAs, the expression images of 33 individual genes were generated by KDE. From the stacked expression images (i.e. the vector field), the representative vectors were selected at the local maxima in neighborhood size 3x3 px squared box. And then the selected vectors were thresholded based on their gene expressions (Figure 3.5A,B). To further remove spurious vectors outside of the tissue region, the vectors were filtered with KNN density (Figure 3.5C). And the whole vector field are normalized with *sctransform* [128].

For *de novo* analysis, the selected vectors were clustered (Figure 3.4A,B), and the bad clusters which mapped to artifacts were removed, and then the similar clusters were merged (Figure 3.6A). There were 30 remaining clusters, and the centroids of the clusters were regarded as signatures of the identified cell types. The *de novo* cell-type signatures determined by SSAM *de novo* mode were consistent with those previously found by segmentation-based analysis and the single-cell RNA sequencing (Figure 3.6C,D) [110]. The signatures were then mapped to the vector field to generate a cell-type map of the *de novo* signatures (Figure 3.4C).

For guided mode analysis, the cell-type signatures from the segmentation-based analysis [110] and the single-cell RNA sequencing data analysis [129] in the previous publication were also mapped to the vector field to generate cell-type maps. Here, for both data, the raw count matrix was normalized with *sctransform* [128], and then the centroids were calculated based on the annotated cell-type information available with the dataset, and the centroids were regarded as the cell-type signature of each data. The resulting cell-type maps were visually similar to each other (Figure 3.6E). Although the resulting cell-type map does not have the information on the number of cells in the image, all three cell-type maps showed more dense cell-like blobs in the image, compared to that of the previous publication. Two clear differences were identified in the resulting cell type maps in terms of the number of these blobs, 1) a lot more Astrocytes expressing *Mfge8* in the SSAM cell-type maps, 2) clearer tissue structures especially in the ventricle area.

Firstly, the abundance of Astrocytes was validated using the marker gene expression, *Mfge8* (Figure 3.4E). The reconstructed KDE image of *Mfge8* shows that the highly expressed regions in the image match well with the regions determined as ‘Astrocyte *Mfge8*’ in the *de novo* cell-

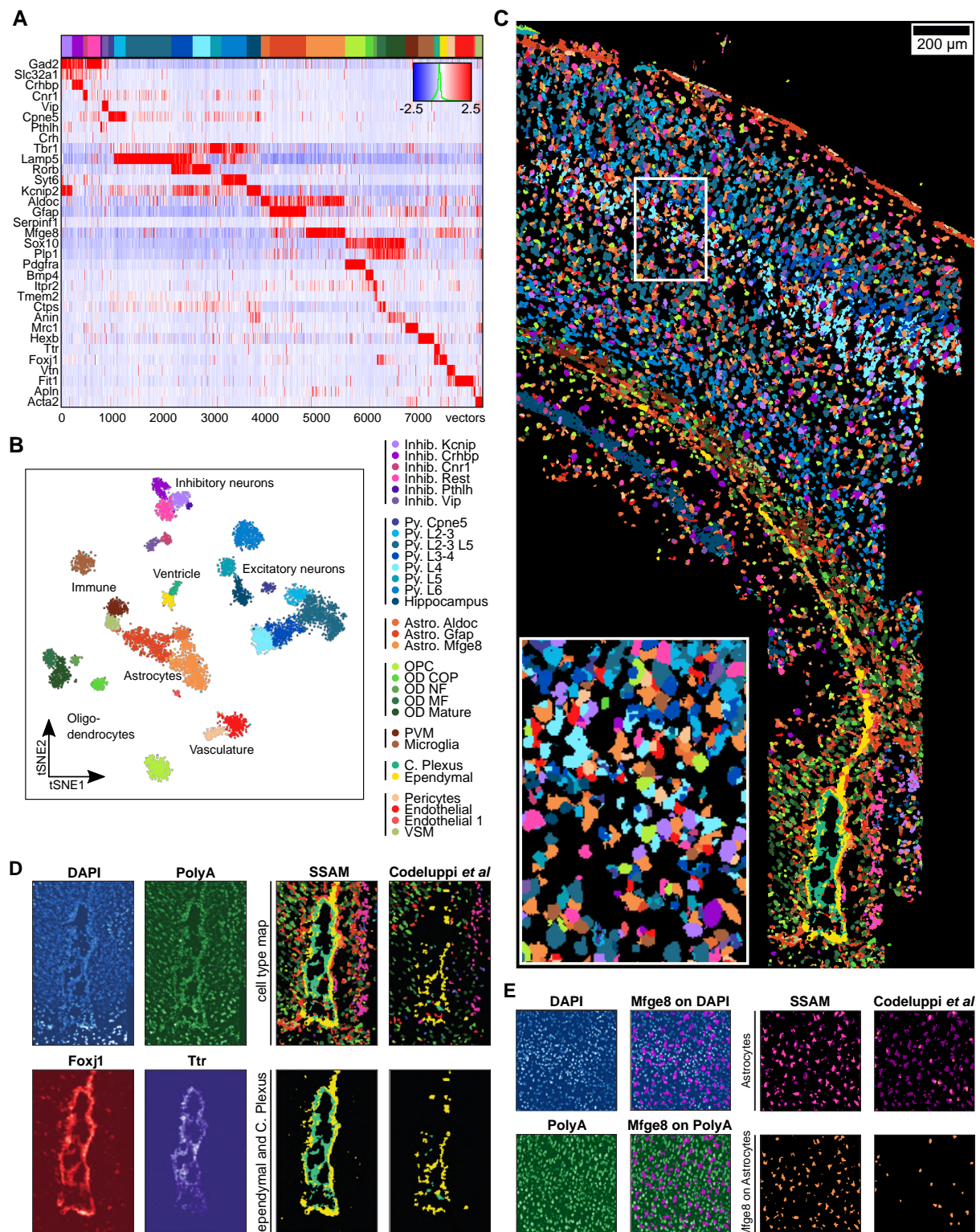


Figure 3.4. SSAM improves astrocyte and ventricle detection in the mouse SSp region.

(A) Gene expression heatmap of the downsampled vectors showing cell-type specific marker gene expression of the vectors within each cluster. The values are z-scored, normalized gene expression. The clustering result is shown on top of the heatmap; (B) A t-SNE map of the downsampled vectors, which shows distinct clusters embedded in 2D space; (C) The *de novo*

cell-type map, showing the spatial organization of cell-types in a spatial context. The zoom panel shows the complex spatial organization of cell types; (D) Validation of the reconstruction ventricle structure by SSAM, DAPI confirms the existence of cells, and each marker gene of two cell types, ependymal (yellow) and choroid plexus (teal), validates the result of SSAM cell-type map; (E) The existence of large population of *Mfge8* expressing astrocytes detected by SSAM was validated with DAPI, PolyA, and the *Mfge8* expression. Both DAPI and *Mfge8* expression confirms the existence of astrocytes. The low PolyA signal, i.e. low total mRNA contents of astrocytes, implies the failure of detection of a lot of astrocytes in the segmentation-based analysis.

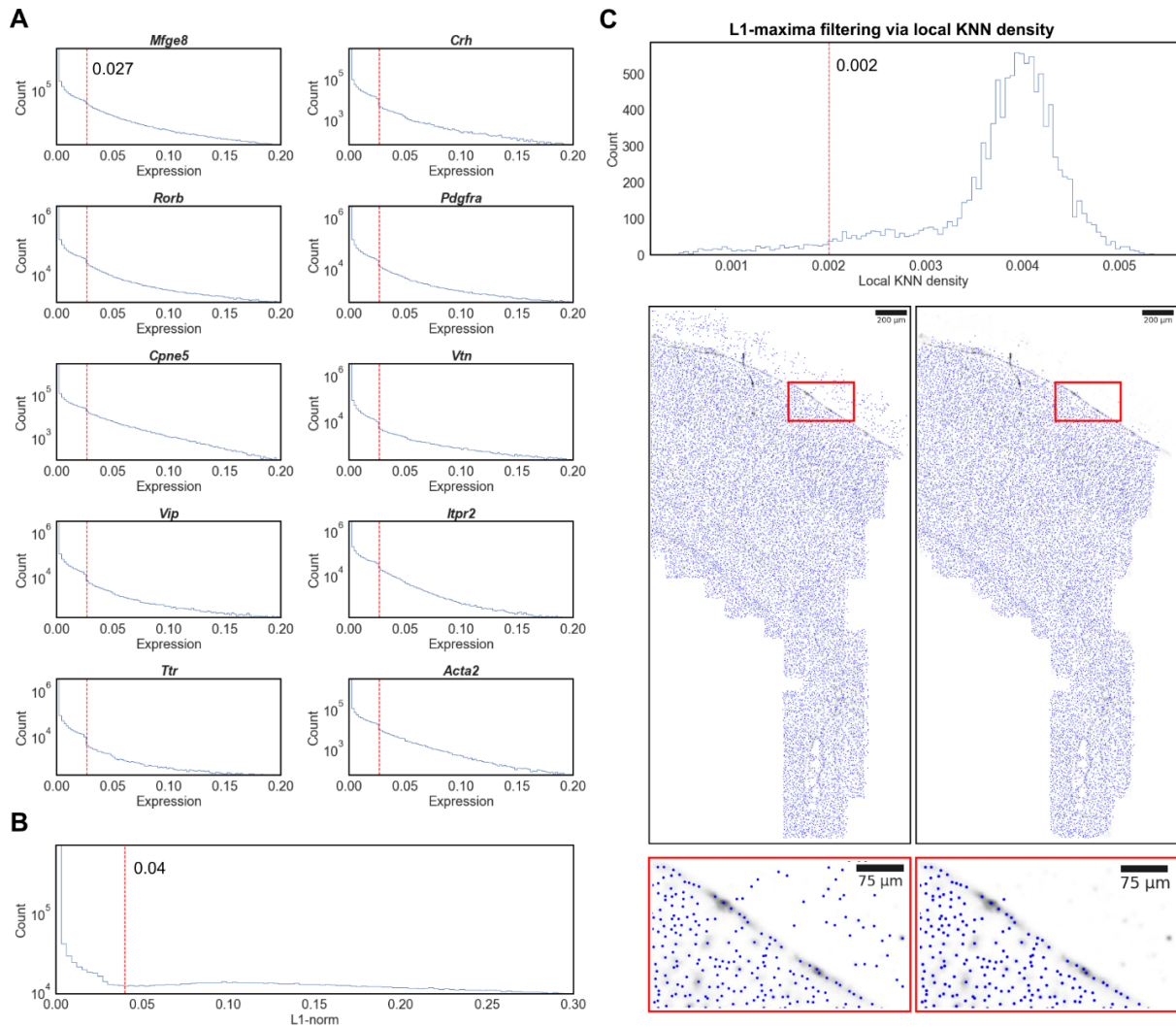


Figure 3.5. Local maxima selection and filtering criteria in the mouse SSp osmFISH dataset.

(A) The gene expression threshold was defined by the location of an observable drop in the gene expression histogram. Here 10 randomly selected genes are presented; (B) The total gene expression threshold was defined based on the total gene expression histogram; (C) Histogram of the local KNN density calculated at all locations of vectors. The threshold of the density was defined based on the histogram to remove the spurious vectors outside of the tissue region.

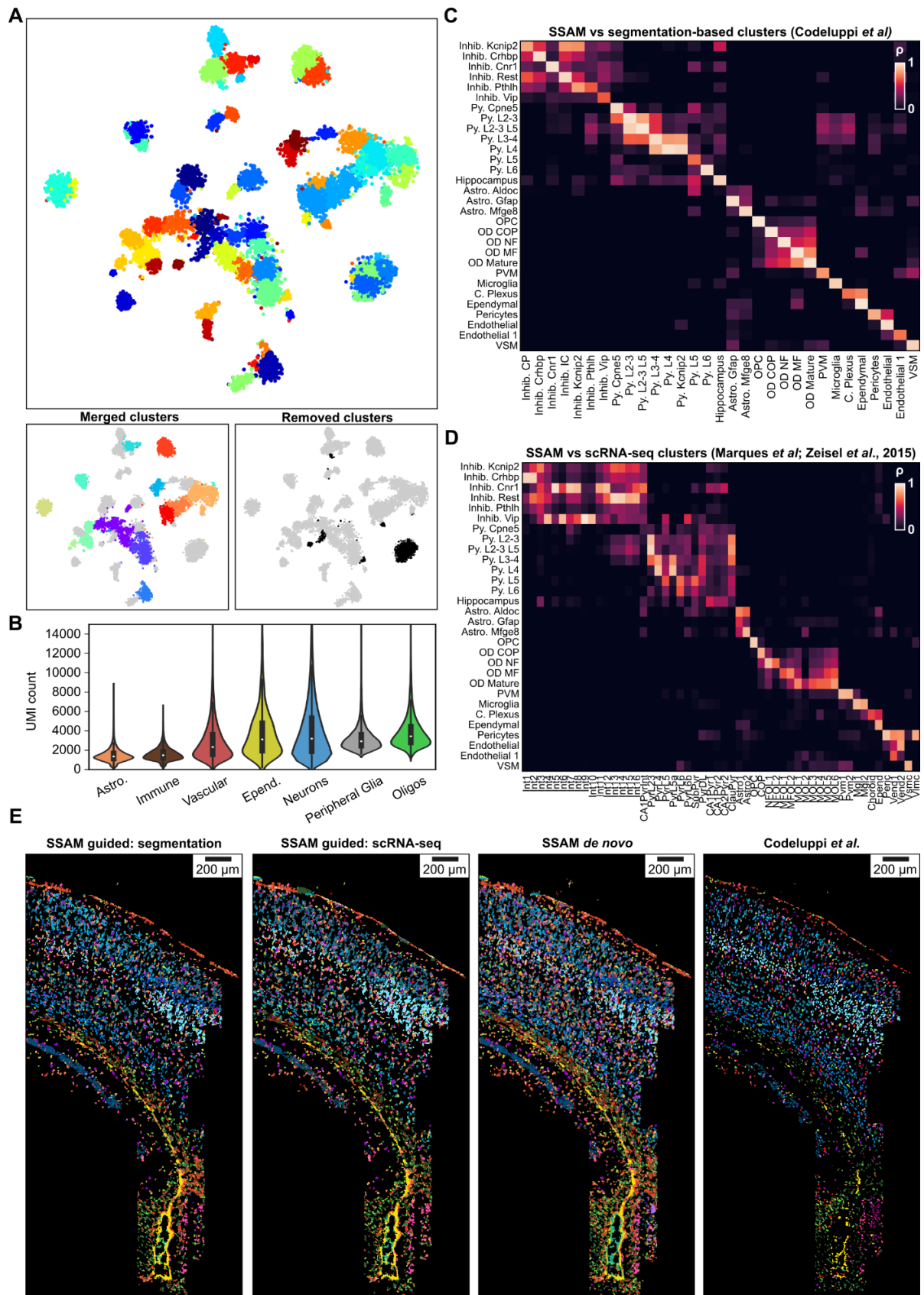


Figure 3.6. Cell-type signature identification and mapping in the mouse SSp osmFISH dataset.

(Note: The analysis described in panel B of this figure was done by Mr. Sebastian Tiesmeyer. The panel B of this figure was also originally produced by Mr. Sebastian Tiesmeyer.)

(A) A 2D t-SNE embedding of clustering local maxima vectors, excluding the ones do not have high correlation to each cluster's medoid (top, see also Figure 3.2B). The merged clusters and the removed clusters are visualized in the same t-SNE map (bottom); (B) Total cell-wise mRNA counts of different cell-type classes in the mouse brain. The violin plot shows UMI counts of 500,000 single mouse brain cells profiled using single-cell RNA sequencing [130], grouped by cell class and ordered by increasing median of UMI counts. Half of the recorded cells from both 'Astrocyte' and 'Immune' classes exhibiting a lower UMI count than the lowest quantile of any other cell class; (C) Comparison of the *de novo* cell-type signatures to the segmentation-based cell-type signatures from Codeluppi *et al.* [110]; (D) Comparison of the *de novo* cell-type signatures to the single-cell RNA sequencing derived cell-type signatures from Marques *et al.* [131] and Zeisel *et al.* [129]; (E) Comparison of cell-type maps generated using SSAM guided and *de novo* mode, (left to right) guided by the cell-type signatures from segmentation-based analysis [110], single-cell RNA sequencing [129,131], SSAM *de novo* cell-type map, and the original figure from Codeluppi *et al.* [110]. The colors of the cell types correspond to the cell-type legend in Figure 3.4.

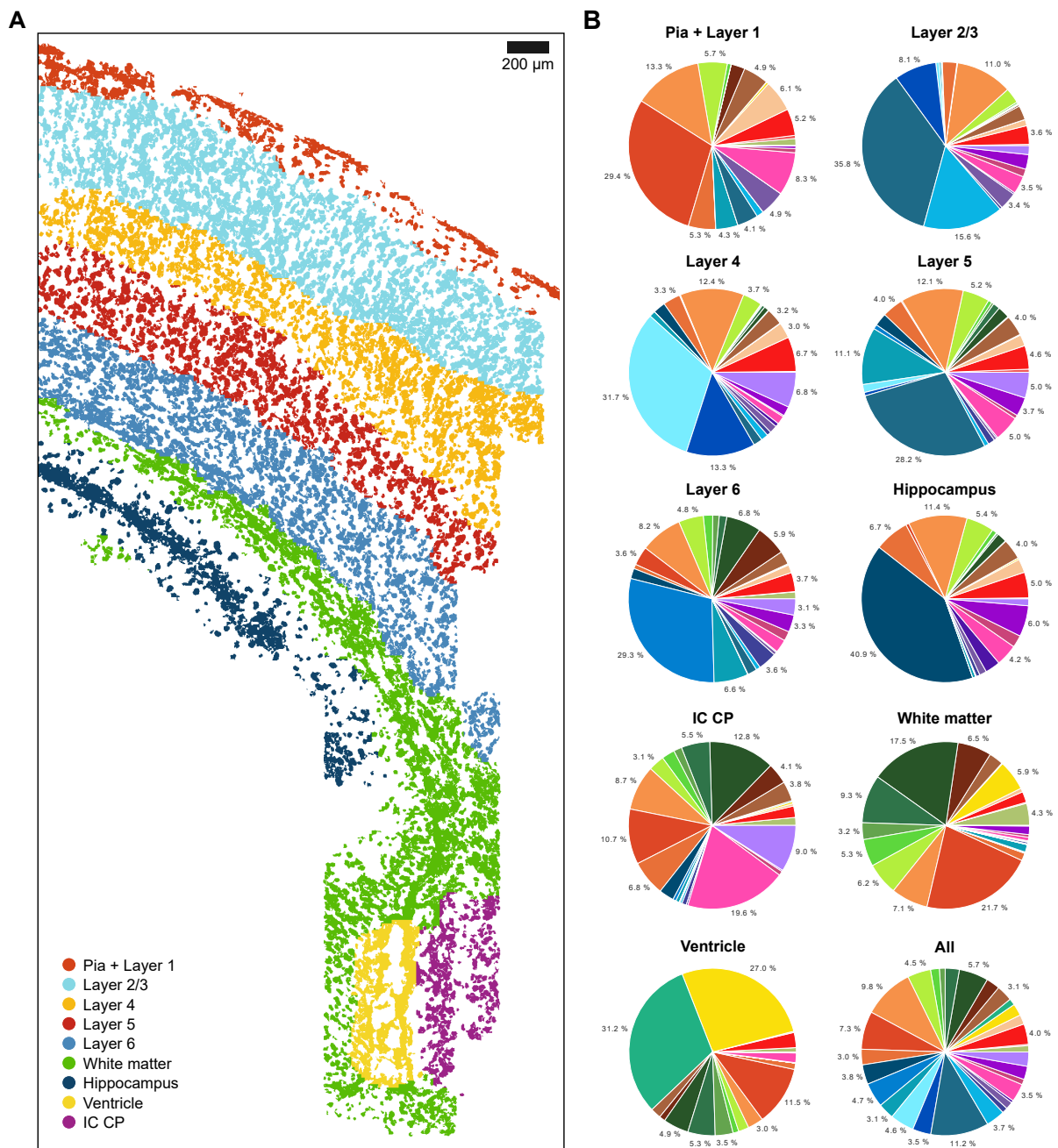


Figure 3.7. SSAM identifies cortical layer tissue domains in the mouse SSp cortex.

(A) Tissue domain map generated by SSAM. Tissue domain signatures were identified from clustering local cell-type composition over sliding 100 μ m circular windows, and projected back onto the cell-type map. The reconstruction shows the various cortical layers (see also Figure 3.3B); (B) Cell-type composition within each tissue domain. The plots show that each domain consists of 7-14% Astrocyte Mfge8 cell types apart from the ventricle, which instead shows a majority of choroid plexus and ependymal cell types. The colors in the pie charts correspond to the cell-type legend in Figure 3.4.

type map. Next, the regions are compared with the DAPI (i.e. the image of stained nuclei) and Poly-A (image of all mRNAs) images. Interestingly, the regions contained at least one nucleus which strongly supports the existence of the cells, whereas the Poly-A signal was weak and not easily distinguished from the background. To explain this discrepancy, it was hypothesized that the total mRNA expression of Astrocytes is lower than the other cell types, which results in low level of signals in the Poly-A image, and therefore leads to the failure of the segmentation algorithm to detect many Astrocytes. This was validated by Mr. Sebastian Tiesmeyer, by comparing the average amount of mRNA counts in the single cells from a single-cell RNA sequencing data [130] (Figure 3.6B, the subfigure was also drawn by Mr. Sebastian Tiesmeyer).

Secondly, the ventricle area was validated with the two marker gene expressions (*Foxj1* and *Ttr*) of two cell types which mainly forms the structure of the ventricle, the ependymal and the choroid plexus cells (Figure 3.4D). The area of high gene expression area in KDE reconstructed image of *Foxj1* and *Ttr* matched well with the Ependymal and choroid plexus cells, respectively. On the other hand, in case of the previous publication, the segmentation algorithm majorly failed to separate single cells within the ventricle region since the cells are very small and tightly packed, as seen in the DAPI and Poly-A images, but instead only identified large segments much bigger than each single cells, which had to be filtered out in the end. Since SSAM does not rely on segmentation algorithms, SSAM reconstructed a much clearer structure of the ventricle compared to the previous result.

Except for the two obvious differences, the resulting cell-type maps were visually similar to the previous result. The visual similarity was validated using a ‘matching score’ (see Method detail section for further details), the ratio of the number of segments matched with the same cell types in the SSAM cell-type map, to the number of the total segments. In other words, the match score is a measure of how well SSAM identified the cells found by the segmentation-based analysis.

By calculating the matching score with SSAM cell-type map guided by the signatures identified in the segmentation-based analysis, most cell types showed a high matching score overall (>0.6 , Table 3.1). The matching score was also calculated with the SSAM cell-type map guided by the signatures identified from single-cell RNA sequencing, and the *de novo* SSAM cell-type map. Since no ground truth of cell types can be found in tissue, the *de novo* signatures identified by SSAM and single-cell RNA sequencing signatures which have a high correlation (>0.8) to the segmentation-based signatures were selected for the comparison. Again the overall matching score was very high (>0.7 , Table 3.2, 3.3). Although there were also several lowly matched cell types (<0.3), the marker gene expression showed better agreement with the cell-type map of the unmatched cell types, compared to the segments in the previous publication (Figure 3.8-3.14). Overall, it was concluded that the cell-type map generated by SSAM is more accurately finds the location of the cell types in spatial context with the osmFISH dataset.

osmFISH cell type	Matched Segments	Total Segments	Matching score
Oligodendrocyte COP	119	171	0.70
Inhibitory CP	66	170	0.39
Pyramidal L5	118	171	0.69
Oligodendrocyte Mature	403	450	0.90
Endothelial	149	252	0.59
pyramidal L4	394	526	0.75
Pyramidal Cpne5	73	97	0.75
Pericytes	22	106	0.21
Vascular Smooth Muscle	4	37	0.11
Inhibitory Crhbp	111	134	0.83
Pyramidal L3-4	117	158	0.74
Oligodendrocyte MF	110	121	0.91
Pyramidal L2-3 L5	302	318	0.95
Pyramidal Kcnip2	16	29	0.55
Astrocyte Mfge8	114	131	0.87
Pyramidal L2-3	195	206	0.95
Inhibitory Cnr1	29	46	0.63
Inhibitory Vip	38	142	0.27
Inhibitory IC	72	91	0.79
Pyramidal L6	410	439	0.93
Hippocampus	93	148	0.63
Ependymal	99	112	0.88
Microglia	52	55	0.95
Oligodendrocyte Precursor cells	117	128	0.91
Inhibitory Kcnip2	49	92	0.53
Oligodendrocyte NF	64	87	0.74
Inhibitory Pthlh	2	104	0.02
Astrocyte Gfap	76	87	0.87
Perivascular Macrophages	1	72	0.01
C. Plexus	8	54	0.15
Endothelial 1	4	105	0.04

Table 3.1. The matching score between the osmFISH PolyA segmented cells vs. SSAM cell-type map guided by the segmentation-based cell-type signatures of osmFISH data.

osmFISH cell type	scRNA-seq cell type	Pearson's r	Matched Segments	Total Segments	Matching score
Oligodendrocyte COP	COP	0.83	128	171	0.75
Endothelial	Vend2	0.8	150	252	0.6
pyramidal L4	S1PyrL5a	0.87	299	526	0.57
Pericytes	Peric	0.83	12	106	0.11
Vascular Smooth Muscle	Vsmc	0.82	4	37	0.11
Inhibitory Crhbp	Int2	0.91	94	134	0.7
Pyramidal L2-3 L5	S1PyrL23	0.89	314	318	0.99
Astrocyte Mfge8	Astro2	0.82	111	131	0.85
Inhibitory Cnr1	Int7	0.83	1	46	0.02
Inhibitory Vip	Int10	0.88	12	142	0.08
Inhibitory IC	Int16	0.94	5	91	0.05
Microglia	Mgl1	0.85	49	55	0.89
Oligodendrocyte Precursor cells	OPC	0.81	118	128	0.92
Oligodendrocyte NF	NFOL1	0.86	58	87	0.67
Perivascular Macrophages	Pvm2	0.91	1	72	0.01

Table 3.2. The matching score between the osmFISH PolyA segmented cells vs. SSAM cell-type map guided by the scRNA-seq cluster centroids for selected cell types based on high gene expression correlation.

osmFISH cell type	<i>de novo</i> cell type	Pearson's r	Matched Segments	Total Segments	Matching score
Oligodendrocyte COP	Oligodendrocyte COP	0.82	119	171	0.70
Inhibitory CP	Inhibitory Kc-nip2	0.96	96	170	0.56
Oligodendrocyte Mature	Oligodendrocyte Mature	0.97	397	450	0.88
Endothelial	Endothelial	0.95	126	252	0.50
pyramidal L4	pyramidal L4	0.80	380	526	0.72
Pyramidal Cpne5	Pyramidal Cpne5	0.95	52	97	0.54
Pericytes	Pericytes	0.97	22	106	0.21
Vascular Smooth Muscle	Vascular Smooth Muscle	0.98	4	37	0.11
Inhibitory Crhbp	Inhibitory Crhbp	0.95	108	134	0.81
Pyramidal L3-4	Pyramidal L3-4	0.97	108	158	0.68
Oligodendrocyte MF	Oligodendrocyte MF	0.96	108	121	0.89
Pyramidal L2-3 L5	Pyramidal L2-3 L5	0.93	305	318	0.96
Astrocyte Mfge8	Astrocyte Mfge8	0.96	113	131	0.86
Pyramidal L2-3	Pyramidal L2-3	0.94	161	206	0.78
Inhibitory Cnr1	Inhibitory Cnr1	0.98	29	46	0.63
Inhibitory IC	Inhibitory Rest	0.98	67	91	0.74
Pyramidal L6	Pyramidal L6	0.97	397	439	0.90
Hippocampus	Hippocampus	0.98	94	148	0.64
Ependymal	Ependymal	0.97	95	112	0.85
Microglia	Microglia	0.93	50	55	0.91

Oligodendrocyte Precursor cells	Oligodendrocyte Precursor cells	0.98	116	128	0.91
Inhibitory Kc-nip2	Inhibitory Pthlh	0.84	26	92	0.28
Oligodendrocyte NF	Oligodendrocyte NF	0.96	63	87	0.72
Astrocyte Gfap	Astrocyte Gfap	0.93	71	87	0.82

Table 3.3. The matching score between the osmFISH PolyA segmented cells vs. SSAM cell-type map guided by the scRNA-seq cluster centroids for selected cell types based on high gene expression correlation.

Caption for figures 3.8-3.14. There were several cell types were found to have a low matching score between the SSAM *de novo* and the original cell-based segmentation cell-type map. (Left panel) red indicates the regions labelled with the cell type in the SSAM cell-type map, blue indicates the cells labelled with this cell type in the cell-based segmentation cell-type map, and white indicates their overlap. (Right panel) the marker gene expression of the corresponding cell type. For all cell types, the marker gene expression better supports the SSAM cell-type calls compared to the segmentation-based cell-type calls.

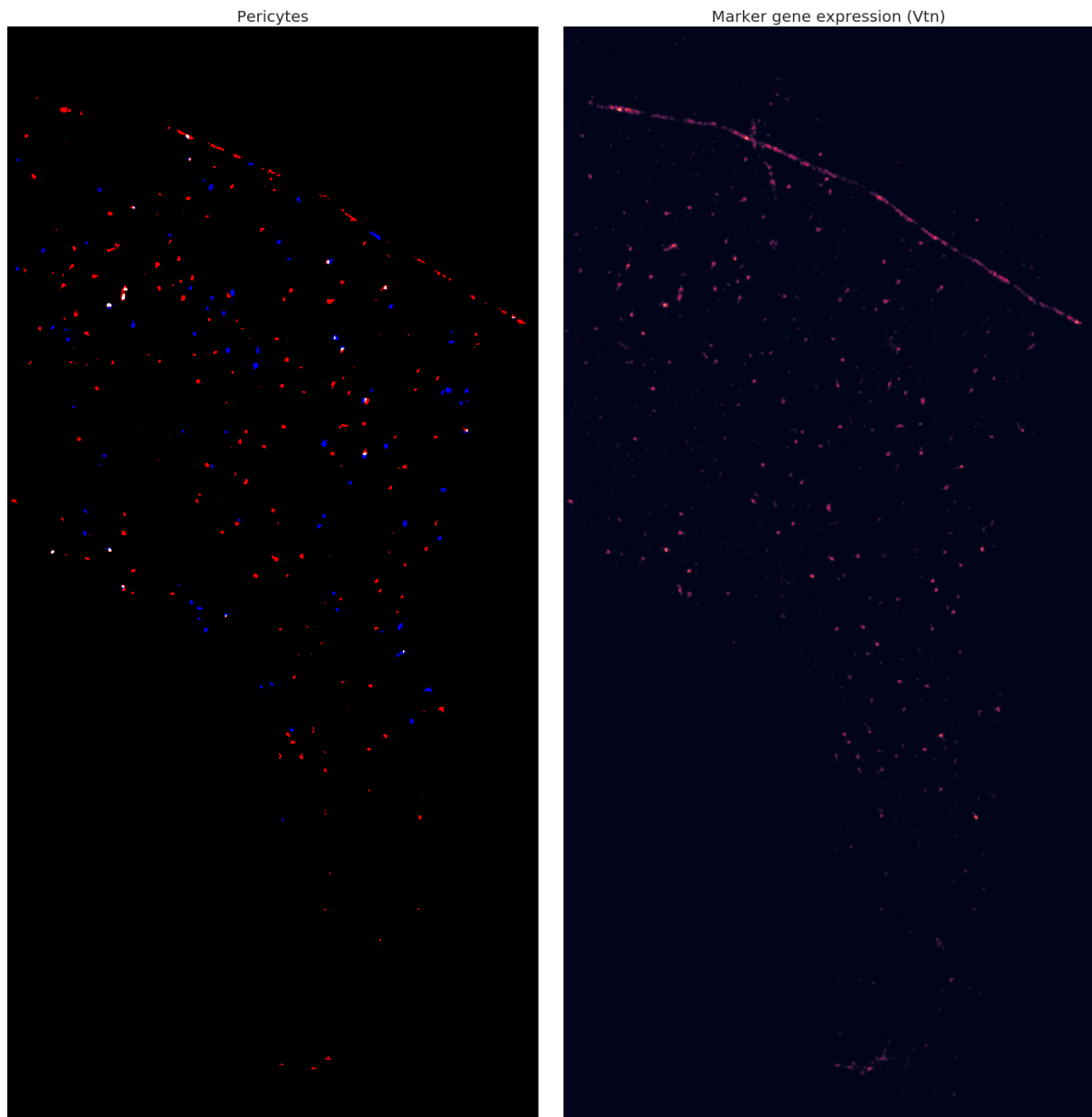


Figure 3.8. Comparison of the cell type (Pericyte) and the corresponding marker gene expression (Vtn) of osmFISH data.

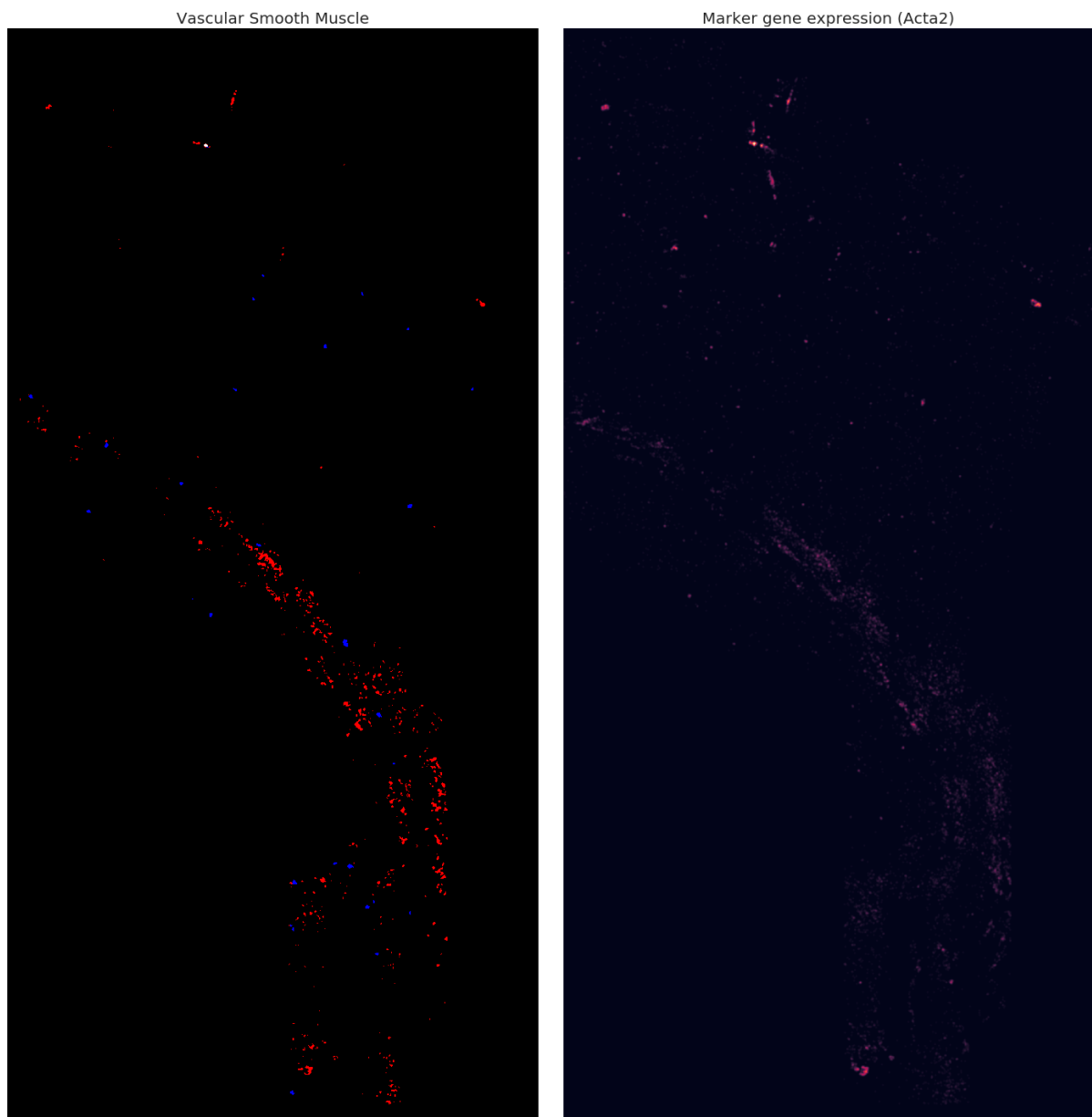


Figure 3.9. Comparison of the cell type (Vascular Smooth Muscle) and the corresponding marker gene expression (Acta2) of osmFISH data.

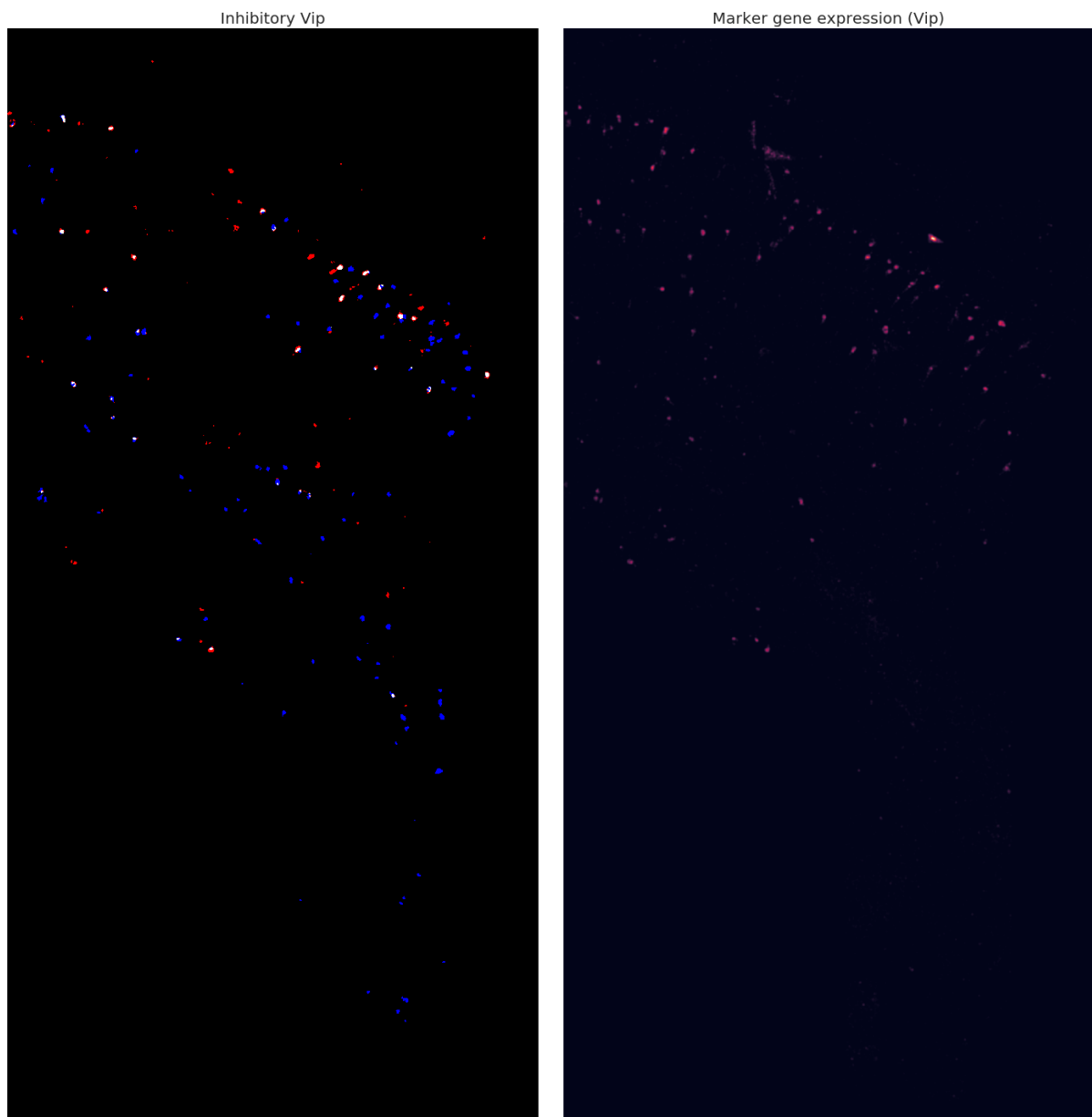


Figure 3.10. Comparison of the cell type (Inhibitory Vip) and the corresponding marker gene expression (Vip) of osmFISH data.

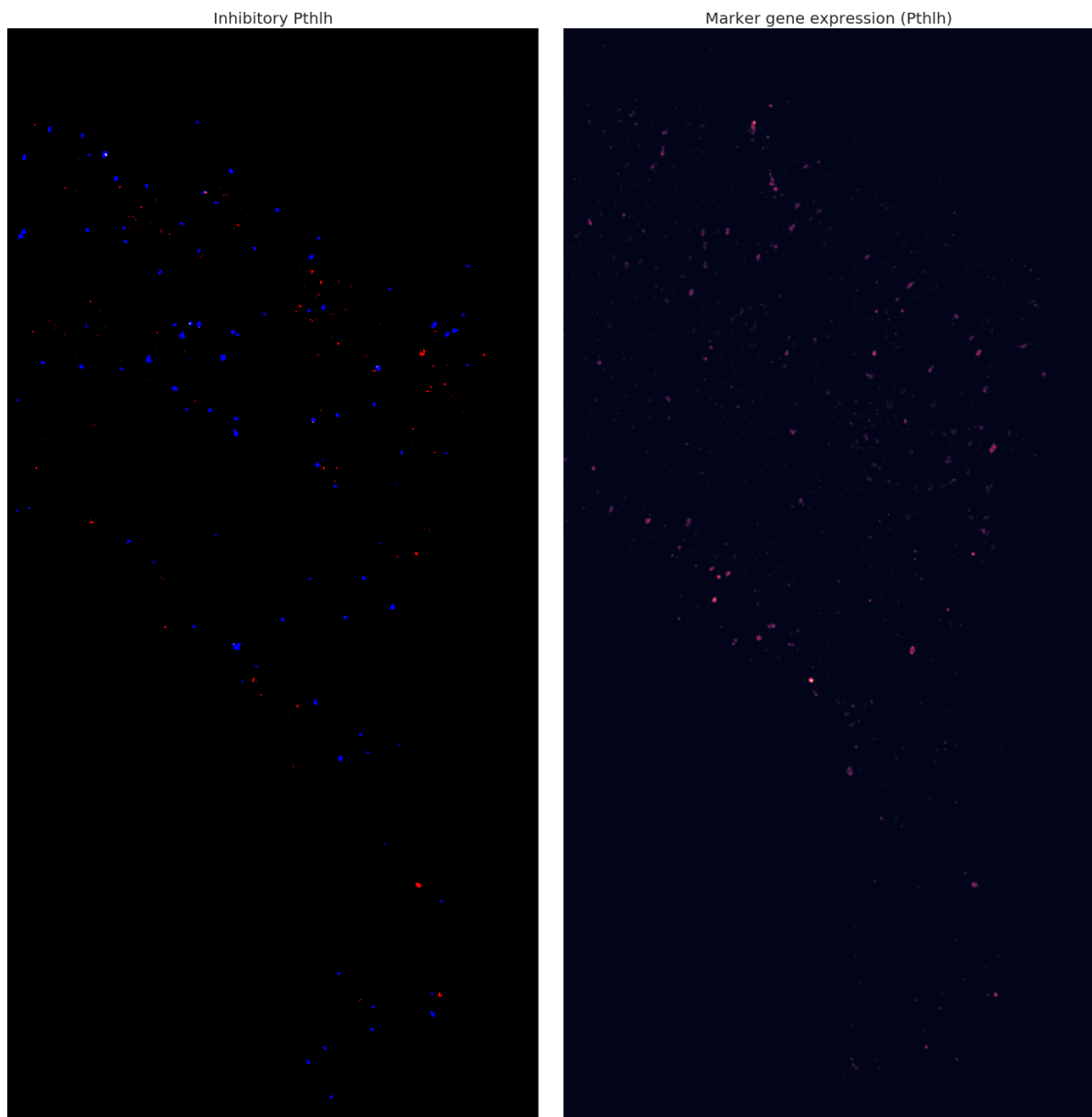


Figure 3.11. Comparison of the cell type (Inhibitory Pthlh) and the corresponding marker gene expression (Pthlh) of osmFISH data.

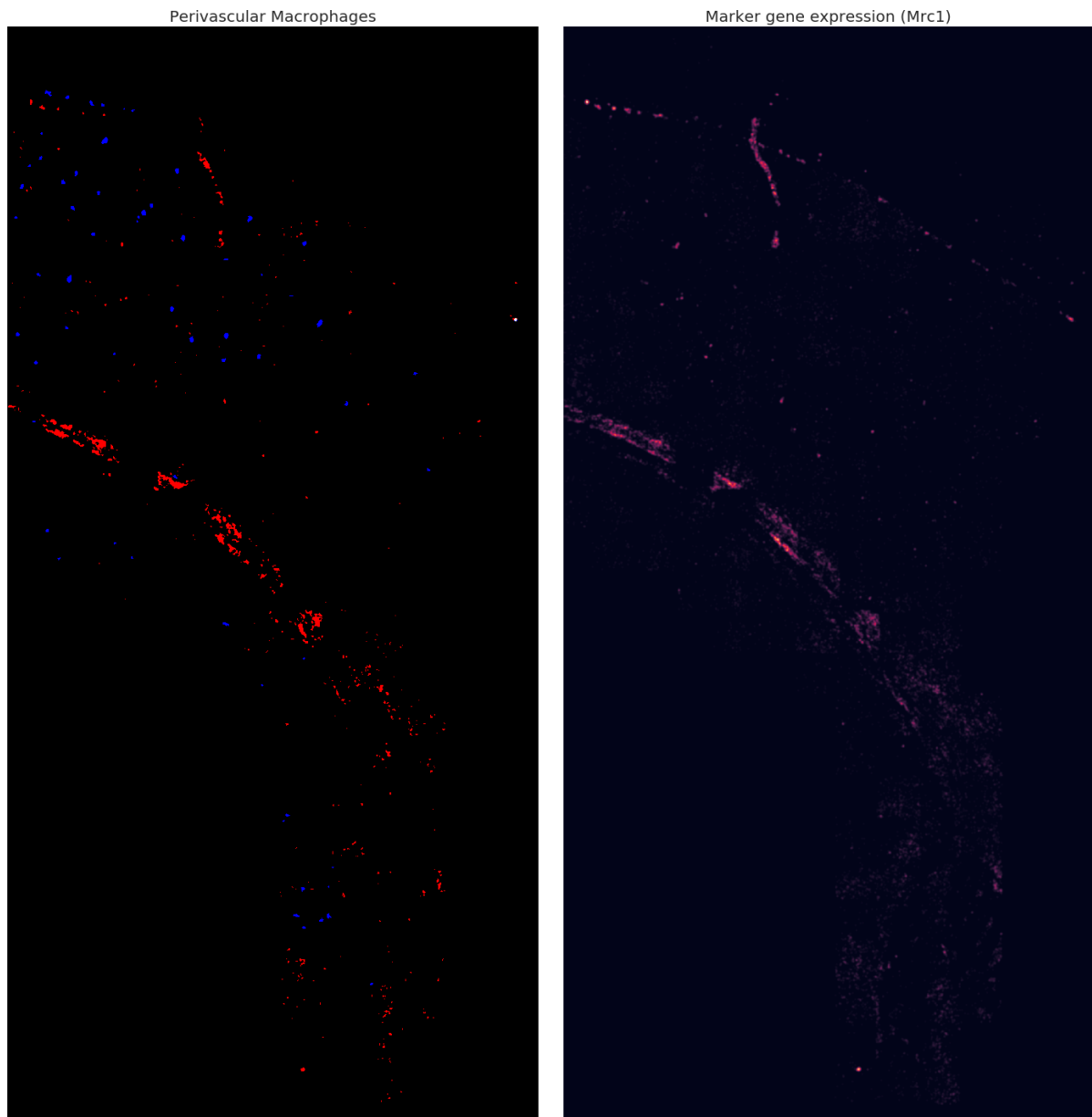


Figure 3.12. Comparison of the cell type (Perivascular Macrophages) and the corresponding marker gene expression (Mrc1) of osmFISH data.

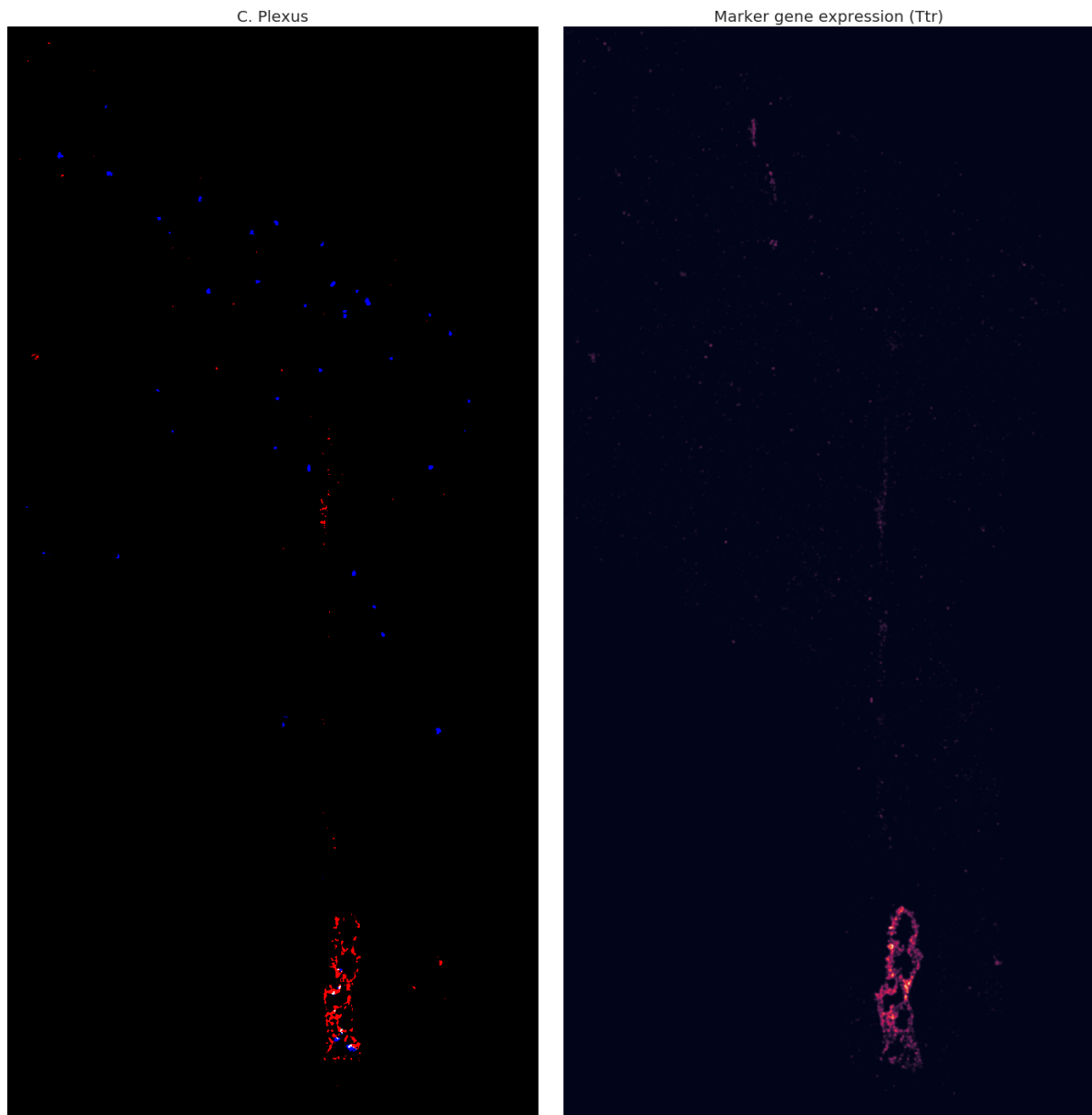


Figure 3.13. Comparison of the cell type (C. Plexus) and the corresponding marker gene expression (Ttr) of osmFISH data.

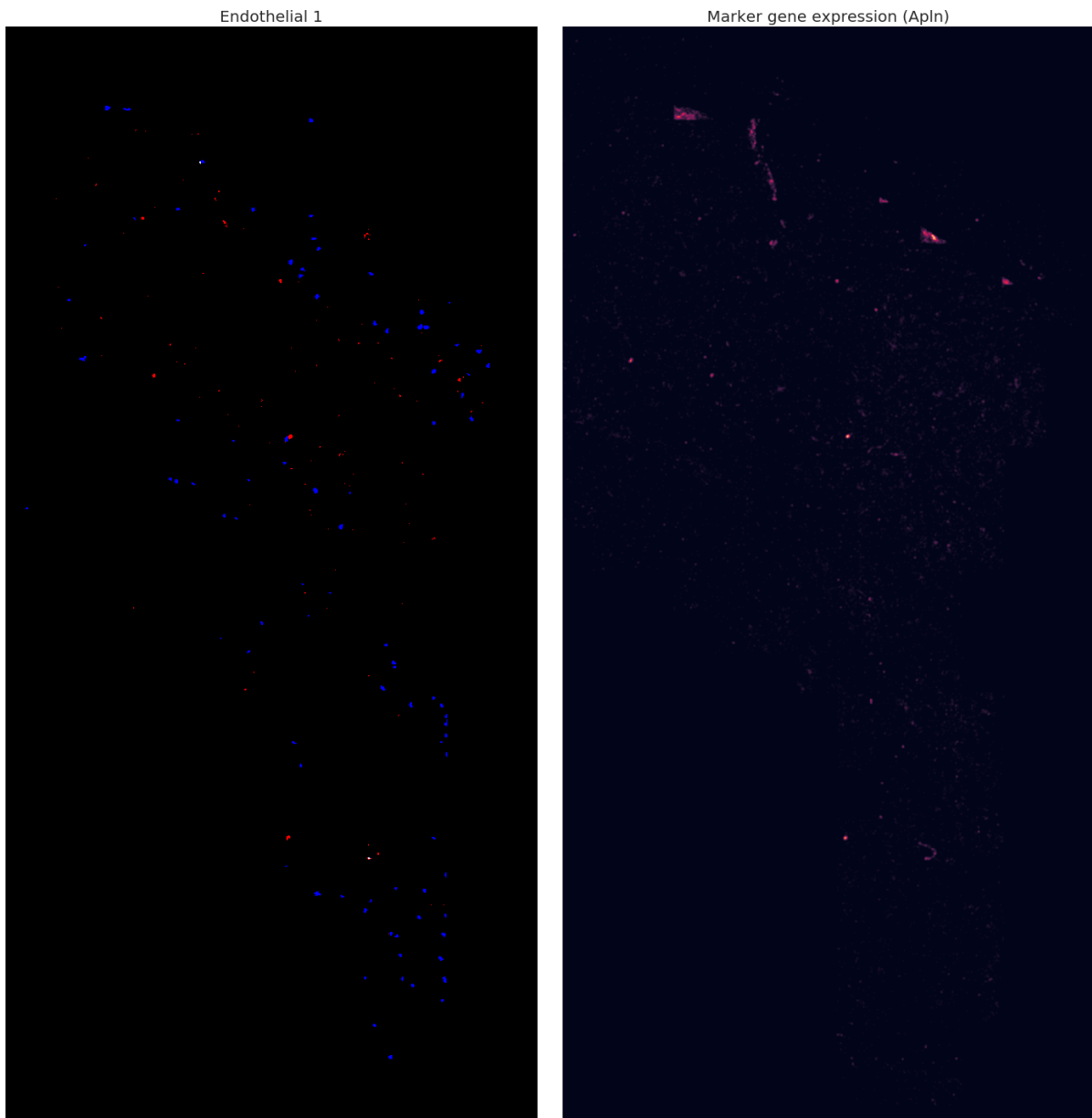


Figure 3.14. Comparison of the cell type (Endothelial 1) and the corresponding marker gene expression (Apln) of osmFISH data.

Based on the high accuracy of the SSAM cell-type maps, the previously known tissue domains in the mouse brain cortex were determined based on the local composition of the cell types in the cell-type map (Figure 3.7). For this, on periodic sparse locations on the cell type map, a circular window was swept and the composition of the pixels cell-types per window was obtained. This results in an image smaller than the cell-type map, with each pixel corresponding to the signatures of the composition of cell types in each window. Then the pixels were clustered using the agglomerative clustering method to annotate the area with similar cell-type compositions. Here the areas which have very similar compositions were manually merged and the ones in the background were removed. Finally, the resulting image was upscaled to the size of the original cell-type map, which represents the area of the tissue domains. The obtained image of the tissue domains was visually similar to that reported in the previous publication, and also matches well with the histological findings on the cortex.

3.2.3 Analysis of mouse hypothalamic preoptic area (POA) imaged by MERFISH

Next, the performance of SSAM was also benchmarked done with the adult mouse hypothalamus preoptic area (POA) data, imaged by MERFISH (Figure 3.15,3.16,3.17) [111]. The data contains the locations of mRNAs of 135 genes obtained in 3D space, from a single layer which consists of 7 sections of the POA region.

Since the location of mRNAs is in 3D space, the analysis was also performed in 3D space, simply extending the dimension of KDE calculation, using a 3D Gaussian kernel. The estimated gene expression values by KDE were obtained at the lattice points of periodic cube 3D lattice, which can be interpreted as a 3D image. Then a 3D vector field was created by combining all of the 3D KDE results of each gene. In other words, each voxel in the 3D lattice carries information of 100 gene expressions at the location. Based on a 3D image of total gene expressions, which was obtained by summing up the gene expression at each voxel, the local maxima of the total gene expression 3D image was selected as the representative vectors. The selected vectors were then thresholded with their gene expressions and the total expressions to further remove spurious vectors (Figure 3.16A,B). Both the resulting selected vectors (Figure 3.16D) and the whole 3D vector field were normalized with *sctransform* [128].

For *de novo* analysis, the selected vectors were then clustered using the same Louvain clustering strategy, as explained in the osmFISH data analysis section (Figure 3.15A,B). Then the clusters which have very similar gene expression signatures were merged, and the bad clusters which were mapped to artifacts were removed using the diagnostic plot (Figure 3.16C). Here, the centroids of the clusters showed a high correlation to those of both the segmentation-based and the single-cell RNA sequencing clusters (figure 3.16E,F). The clusters were then regarded

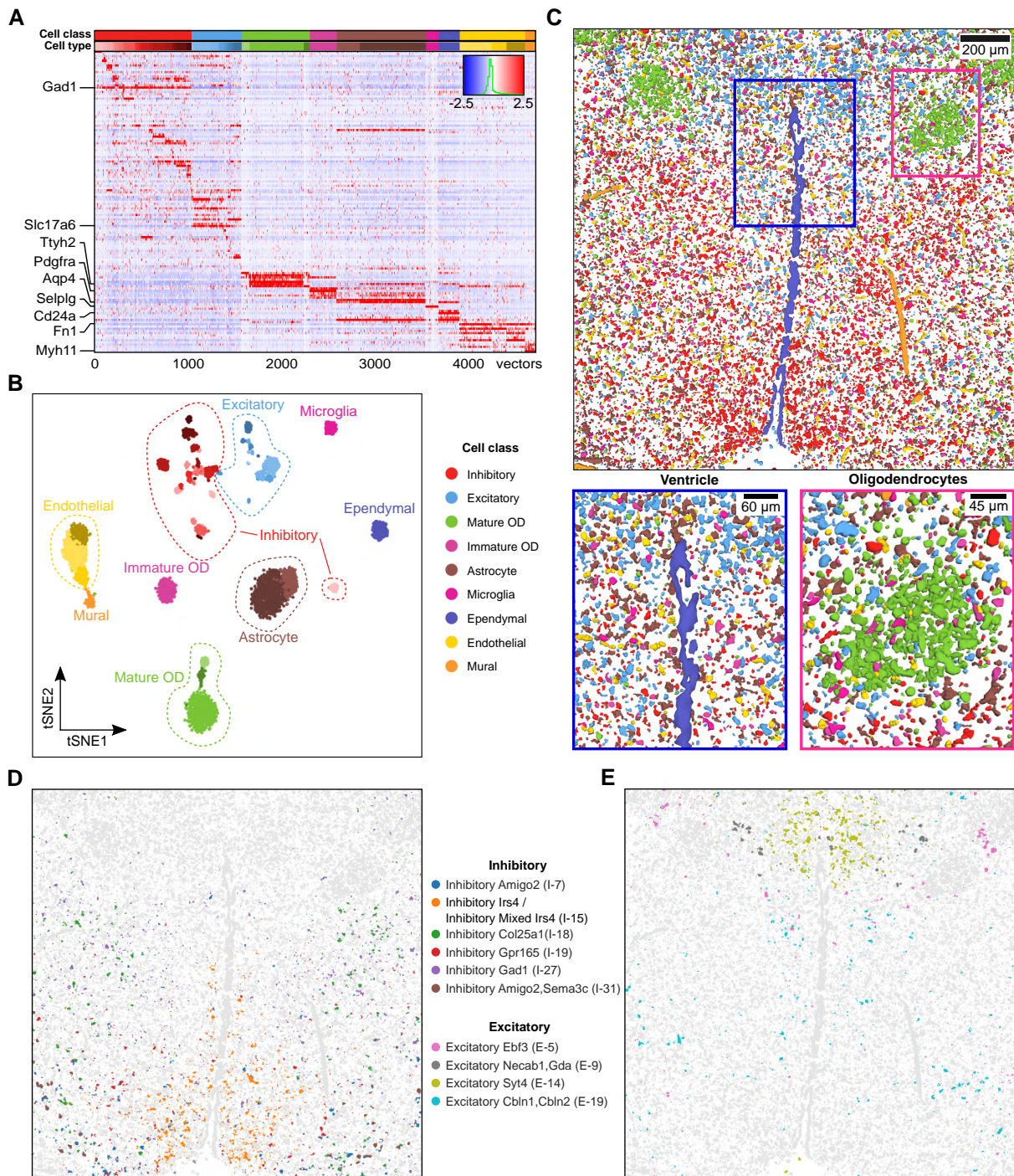


Figure 3.15. SSAM 3D cell-type map confirms rich diversity of heterogeneous cells in the posterior hypothalamic POA.

(A) Gene expression heatmap of the downsampled vectors showing cell-type specific marker gene expression of the vectors within each cluster. The values are z-scored, normalized gene expression. The clustering result (cell type, lower bar) and their cell class (upper bar) are shown on top of the heatmap; (B) A t-SNE map of the downsampled vectors, which shows distinct clusters embedded in 2D space; (C) The *de novo* 3D cell-type map, showing spatial organization of cell-types in spatial context. The zoom panels show the complex spatial organization of cell

types in the ventricle region, and the spatial cluster of oligodendrocytes; (D) spatial localization of various inhibitory cell-type signatures. We found a number of inhibitory cell types which both matched expression signature and tissue localization described by Moffitt *et al.* [111]; (E) As panel D, but for excitatory cell types.

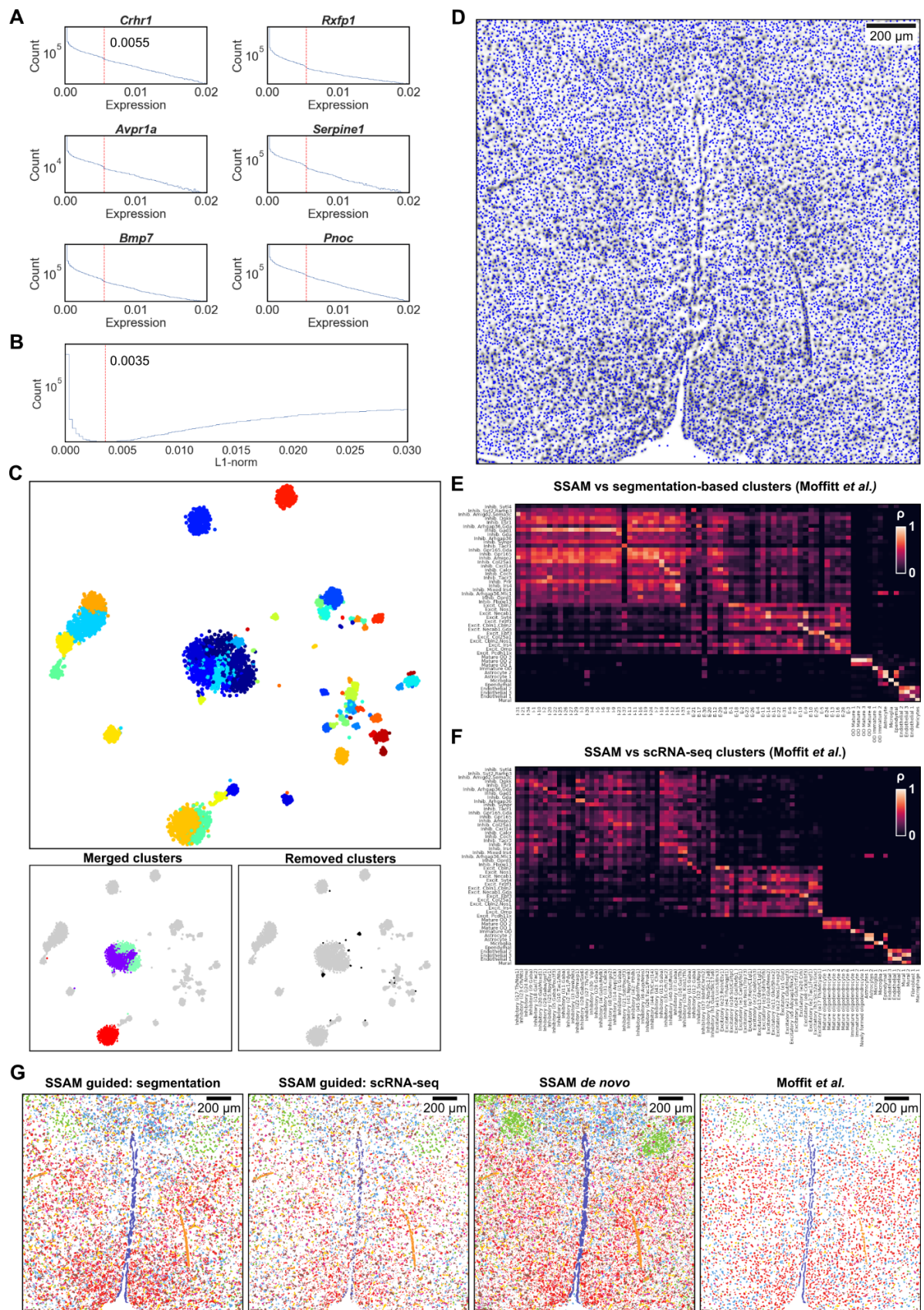


Figure 3.16. Local maxima selection, cell-type signature identification, and mapping in the mouse POA MERFISH dataset.

(A) The gene expression threshold (red vertical line) was defined by the location of an observable drop in the gene expression histogram. Here 6 randomly selected genes are presented; (B) The total gene expression (red vertical line) threshold was defined based on the total gene expression histogram; (C) A 2D t-SNE embedding of clustering local maxima vectors, excluding the ones do not have high correlation to each cluster's medoid (top, see also Figure 3.2B). The merged clusters and the removed clusters are visualized in the same t-SNE map (bottom); (D) Selected local maxima on the vector field.; (E) Comparison of the SSAM *de novo* cell-type signatures to the segmentation-based cell-type signatures from Moffitt *et al.* [111]; (F) Comparison of the SSAM *de novo* cell-type signatures to the scRNA-seq derived cell-type signatures from Moffitt *et al.* [111]; (G) Comparison of cell-type maps generated using SSAM guided and *de novo* mode, (left to right) guided by the cell-type signatures from segmentation-based analysis and single-cell RNA sequencing [111], SSAM *de novo* cell-type map, and the original figure from Moffitt *et al.* [111]. The colors of the cell classes correspond to the cell-class legend in Figure 3.15.

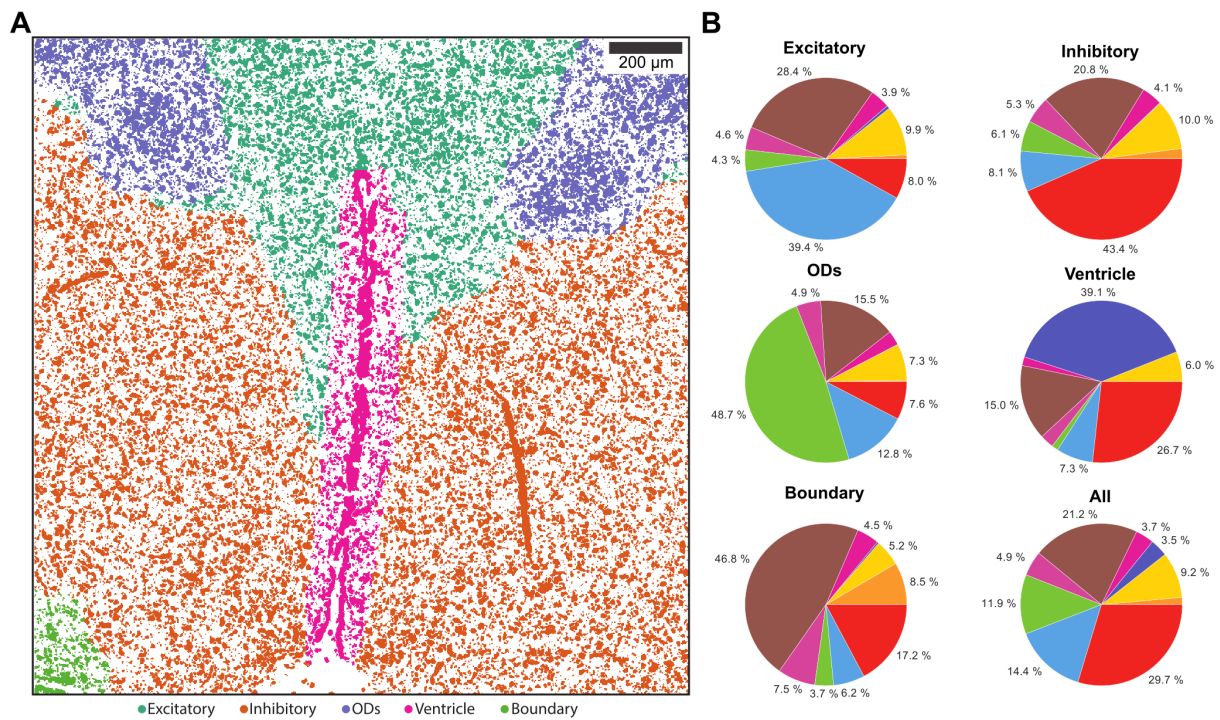


Figure 3.17. SSAM identifies enriched inhibitory and excitatory tissue domains in the posterior hypothalamic POA.

(A) Tissue domain map generated by SSAM. Tissue domain signatures were identified from clustering local cell-type composition over sliding 100 μ m circular windows and projected back onto the cell-type map. The reconstruction shows the various cortical layers (see also Figure 3.3B); (B) Cell-type composition within each tissue domain. The plots show the composition ratio of approximately 5:1 of inhibitory to excitatory cell types in the inhibitory tissue domain and vice versa. The colors in the pie charts correspond to the cell-class legend in Figure 3.15B.

as the cell-type signatures in the tissue, each mapped to the vector field, and merged into the final 3D *de novo* cell-type map. For guided analysis, the cell-type signatures obtained from the previous publication [111], from the segmentation-based analysis and the single-cell RNA sequencing, were mapped to the vector field to create guided cell-type maps.

To visualize the resulting 3D cell-type map, two different strategies were used. First, simply slice the image in the middle of the z-axis. With this strategy, three movies were generated by sweeping z-axis by 1 μm , which shows the 3D structure of all cell-types, the neuronal cell-types, and the astrocytes (the movies can be found bioRxiv online, doi:10.1101/800748). Secondly, the 3D cell-type map was volumetrically rendered in 3D space, by determining the surface boundaries between the continuous blobs in the cell-type map. For example, the resulting 3D rendered *de novo* cell-type map is shown in Figure 3.15C (see Method detail section for further details).

The resulting cell-type maps were visually similar to the one in the previous publication (Figure 3.16G) [111]. This visual similarity between the cell-type maps was validated using the matching score. The matching score was calculated between the SSAM cell-type maps (two guided, and one *de novo*) and the cell-type annotated segments in the previous publication (Table 3.4, 3.5, 3.6), using the segmentation-based signatures, and both the single-cell RNA sequencing based signatures and the SSAM *de novo* signatures having high correlation to the segmentation-based signatures. The matching score was overall high (>0.8), which validates the visual similarity between the SSAM cell-type maps and the cell-type map from the previous segmentation-based analysis.

For further comparison to validate the accuracy of *de novo* cell-type map, the neuronal *de novo* clusters which show a high correlation to the segmentation-based clusters were selected and compared with the previous finding [111]. The 7 inhibitory and 4 excitatory neuronal cell types showed similar spatial patterns as previously discovered (Figure 3.15D,E, 3.18).

Despite the similarity between the cell-type maps, there was also one big difference, which is the abundance of astrocytes. In *de novo* cell-type map, SSAM detected many blobs of astrocytes that were not identified as astrocytes in previous publication [111], and some of them even couldn't be identified as any cell in the original cell-type map. Such astrocytes were validated with the marker gene expression (*Aldh1l1*), which shows the existence of the astrocyte cells, also shows the potential accuracy of the SSAM cell-type map in terms of finding the spatial organization of the cell types (Figure 3.19).

Based on the SSAM cell-type maps, the tissue domains were identified by sweeping a 3D spherical window in the 3D cell-type map (Figure 3.17). The resulting 3D domain map was visualized at the center of the z-axis (4 μm), which shows the tissue structure of the hypothalamus preoptic region.

MERFISH cell type	Matched segments	Total segments	Matching score
Astrocyte	720	744	0.97
Endothelial 1	282	301	0.94
Endothelial 2	17	19	0.89
Endothelial 3	110	117	0.94
Ependymal	219	219	1
Microglia	97	106	0.92
OD Immature 1	217	223	0.97
OD Immature 2	2	7	0.29
OD Mature 1	24	49	0.49
OD Mature 2	213	241	0.88
OD Mature 3	6	6	1
OD Mature 4	6	10	0.6
Pericytes	60	62	0.97
I-1	0	352	0
I-2	74	120	0.62
I-3	17	47	0.36
I-4	13	17	0.76
I-5	4	5	0.8
I-6	3	13	0.23
I-7	126	245	0.51
I-8	26	31	0.84
I-9	29	50	0.58
I-10	14	27	0.52
I-11	61	107	0.57
I-12	90	139	0.65
I-13	191	316	0.6
I-14	70	89	0.79
I-15	168	174	0.97
I-16	70	99	0.71
I-18	107	135	0.79
I-19	4	11	0.36
I-20	9	10	0.9
I-21	3	24	0.12
I-22	7	10	0.7

I-23	65	75	0.87
I-24	11	15	0.73
I-25	24	44	0.55
I-26	1	1	1
I-27	12	12	1
I-29	5	10	0.5
I-30	4	6	0.67
I-31	19	23	0.83
I-33	2	3	0.67
I-34	5	5	1
I-37	1	1	1
H-1	1	1	1
E-1	14	15	0.93
E-2	3	6	0.5
E-3	14	18	0.78
E-4	24	35	0.69
E-5	1	1	1
E-6	9	24	0.38
E-7	35	79	0.44
E-8	12	18	0.67
E-9	19	19	1
E-10	6	14	0.43
E-11	7	20	0.35
E-12	5	35	0.14
E-13	90	117	0.77
E-14	239	298	0.8
E-15	42	56	0.75
E-16	123	145	0.85
E-17	24	29	0.83
E-18	64	123	0.52
E-19	48	50	0.96
E-20	6	8	0.75
E-21	5	6	0.83
E-22	0	5	0
E-23	23	27	0.85

E-24	26	41	0.63
E-25	1	3	0.33
E-26	1	1	1
E-28	19	19	1
E-29	1	1	1
E-30	1	1	1
E-31	2	2	1

Table 3.4. The matching score between the MERFISH segmented cells vs. SSAM cell-type map guided by the segmentation-based cell-type signature of MERFISH data.

MERFISH cell type	scRNA-seq cell type	Pearson' s r	Matched segments	Total seg-ments	Matching score
Astrocyte	Astrocytes 2	0.91	516	744	0.69
OD Immature 1	Immature oligodendrocyte 1	0.91	220	223	0.99
Microglia	Microglia 2	0.9	101	106	0.95
Endothelial 3	Mural 2	0.86	75	117	0.64
E-5	Excitatory (e8:Cck/Ebf3)	0.9	1	1	1

Table 3.5. The matching score between the MERFISH segmented cells vs. SSAM cell-type map guided by scRNA-seq cluster centroids for selected cell types based on high gene expression correlation.

MERFISH cell type	<i>de novo</i> cell type	Pearson's r	Matched segments	Total segments	Matching score
Astrocyte	Astrocyte 1	0.94	739	744	0.99
	Astrocyte 2	0.89			
Microglia	Microglia	0.98	101	106	0.95
Ependymal	Ependymal	0.91	219	219	1
Endothelial 1	Endothelial 1	0.93	295	301	0.98
	Endothelial 2	0.87			
Endothelial 3	Endothelial 3	0.97	112	117	0.95
Pericytes	Mural	0.95	60	62	0.96
OD Immature 1	Immature OD	0.92	223	223	1
OD Mature 2	Mature OD 2	0.97	241	241	1
I-7	Inhibitory Amigo2	0.84	127	245	0.51
I-15	Inhibitory Irs4	0.87	166	174	0.95
	Inhibitory Mixed Irs4	0.80			
I-18	Inhibitory Col25a1	0.88	69	135	0.51
I-19	Inhibitory Gpr165	0.80	6	11	0.54
I-27	Inhibitory Gad1	0.92	9	12	0.75
I-31	Inhibitory Amigo2,Sema3c	0.91	23	23	1
E-5	Excitatory Ebf3	0.86	1	1	1
E-9	Excitatory Necab1,Gda	0.92	17	19	0.89
E-14	Excitatory Syt4	0.87	169	298	0.56
E-19	Excitatory Cbln1,Cbln2	0.84	48	50	0.96

Table 3.6. The matching score between the MERFISH segmented cells vs. SSAM cell-type map guided by scRNA-seq cluster centroids for selected cell types based on high gene expression correlation.

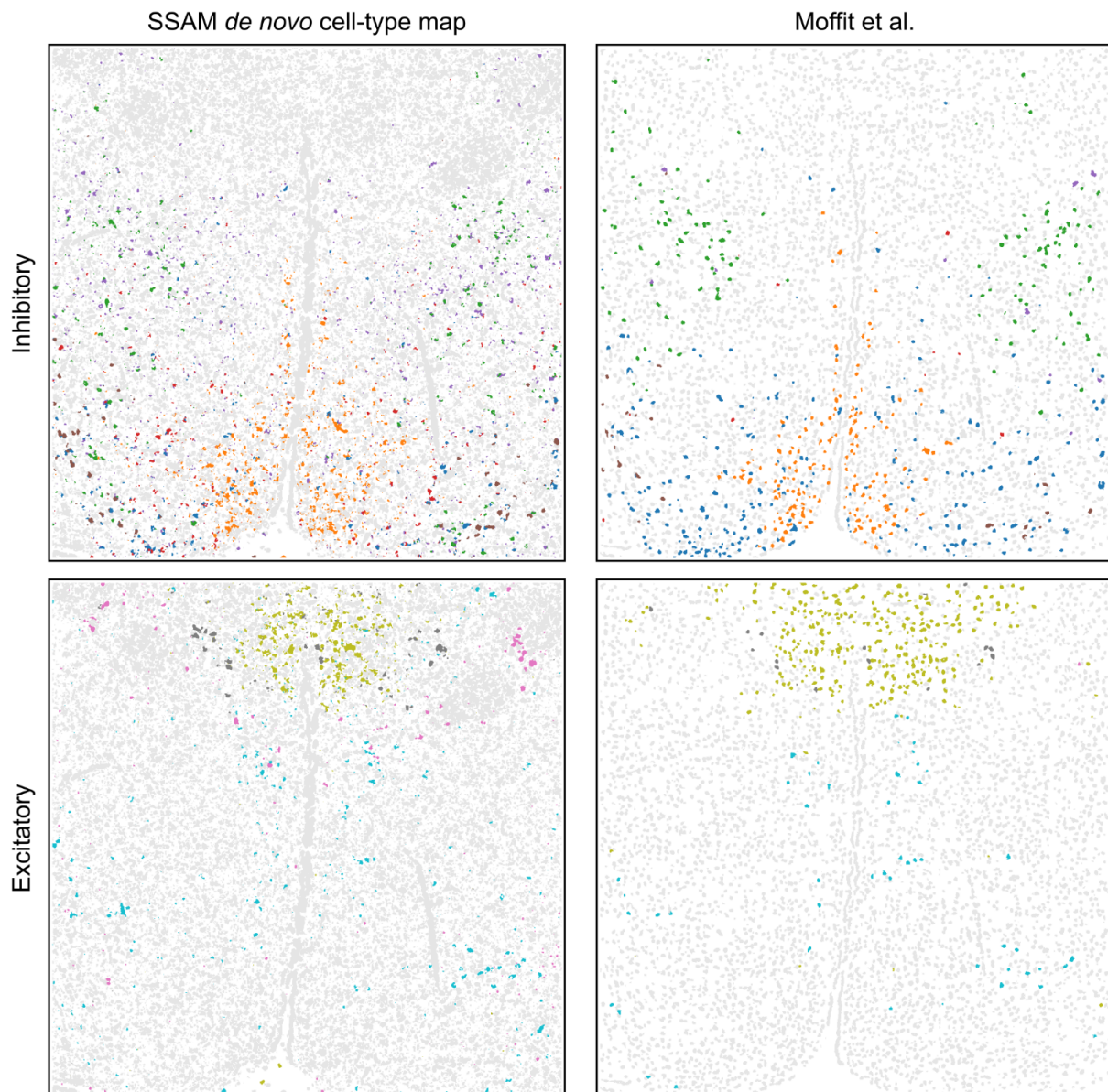


Figure 3.18. Side-by-side comparison of Astrocytes identified by SSAM guided mode vs gene expression of Aldh111, in MERFISH dataset.

A number of inhibitory (top row) and excitatory (bottom row) cell types were found to have similar tissue localization patterns between the SSAM *de novo* cell-type map (left column) and the segmentation-based cell-type map (right column). The cell-class legend can be found in Figure 3.15B.

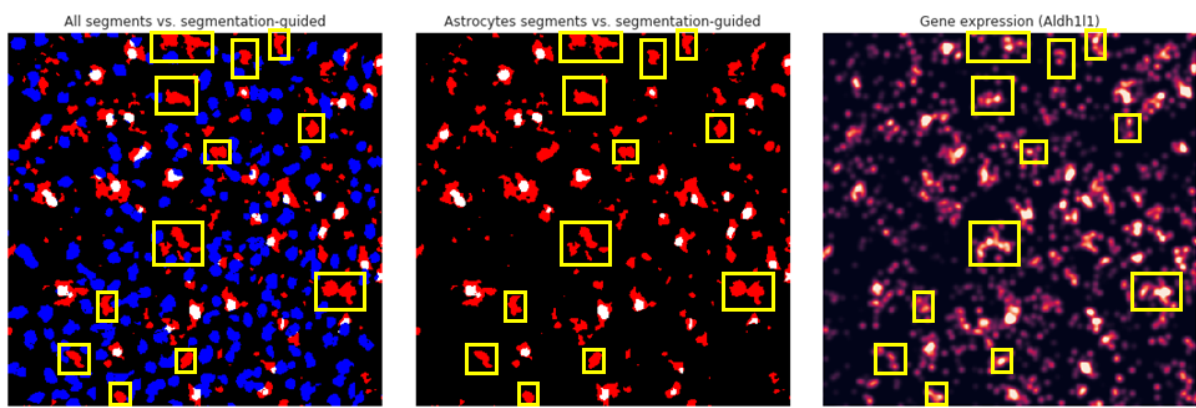


Figure 3.19. Side-by-side comparison of Astrocytes identified by SSAM guided mode vs gene expression of Aldh111, in MERFISH dataset.

(Left) SSAM guided-mode cell type is colored in red, the segments of all cell types by Moffit et al. are colored in blue, and the overlap is colored in white. (Middle) The same as the left panel, but without the segments assigned to other than Astrocytes. (Right) Gene expression of Aldh111. Although Aldh111 is not a specific marker for Astrocytes, it is one of the genes highly expressed in Astrocytes, which confirms the existence of Astrocytes uniquely identified by SSAM (highlighted in yellow).

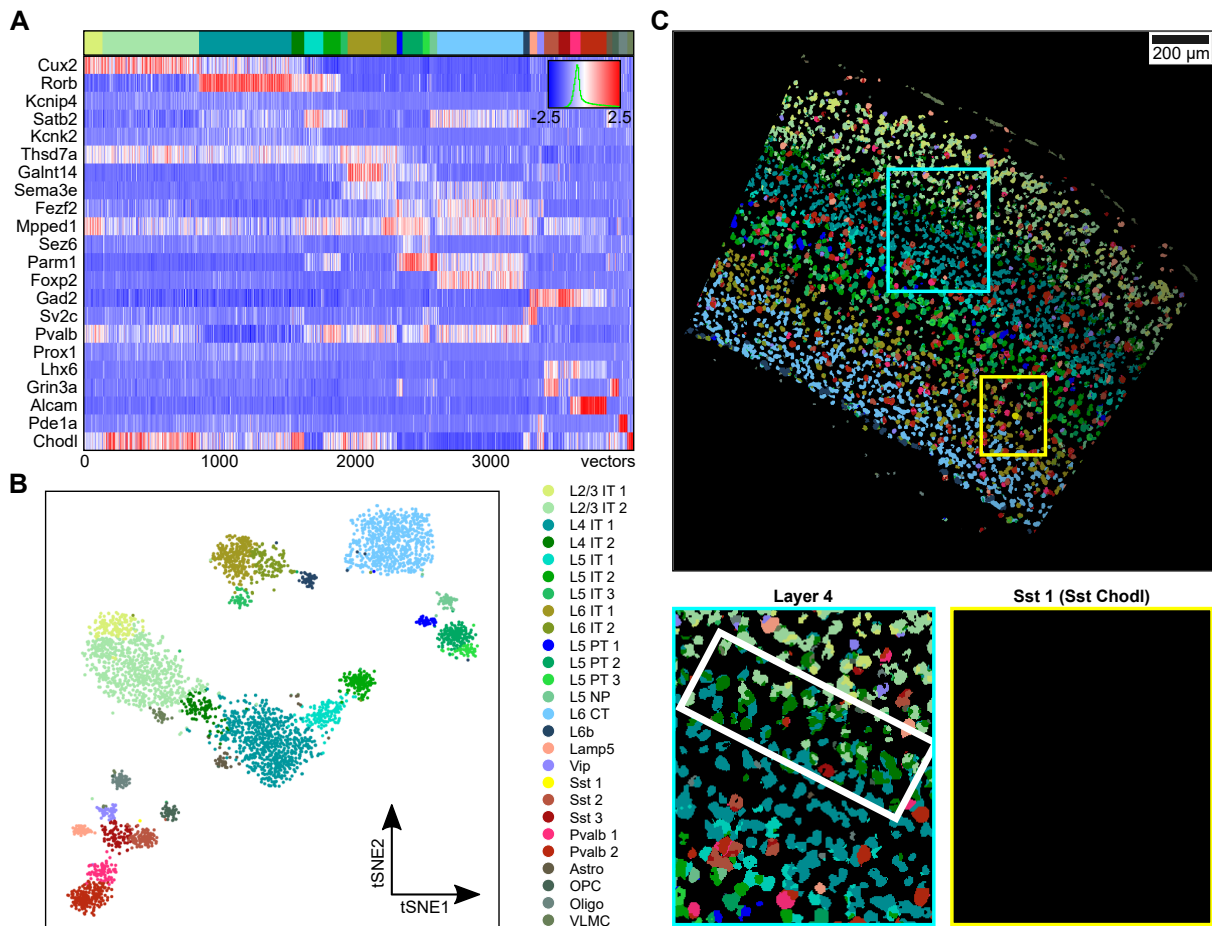


Figure 3.20. SSAM identifies a new cell type in L4 and confirms rare Sst Chodl cell type in the mouse VISp region.

(A) Gene expression heatmap of the downsampled vectors showing cell-type specific marker gene expression of the vectors within each cluster. The values are z-scored, normalized gene expression of the vectors within each cluster. The clustering result is shown on top of the heatmap, coloring based on the highest correlating single-cell RNA sequencing based cell-type signature [132]; (B) A t-SNE map of the downsampled vectors, which shows distinct clusters embedded in 2D space; (C) The *de novo* cell-type map, showing spatial organization of cell-types in spatial context. Lower images zoom in on the highlighted tissue regions of the new cell type found in the L4 superficial region (boxed in white) and rare Sst Chodl cell type. The colors of the cell types correspond to the cell-type legend in panel B.

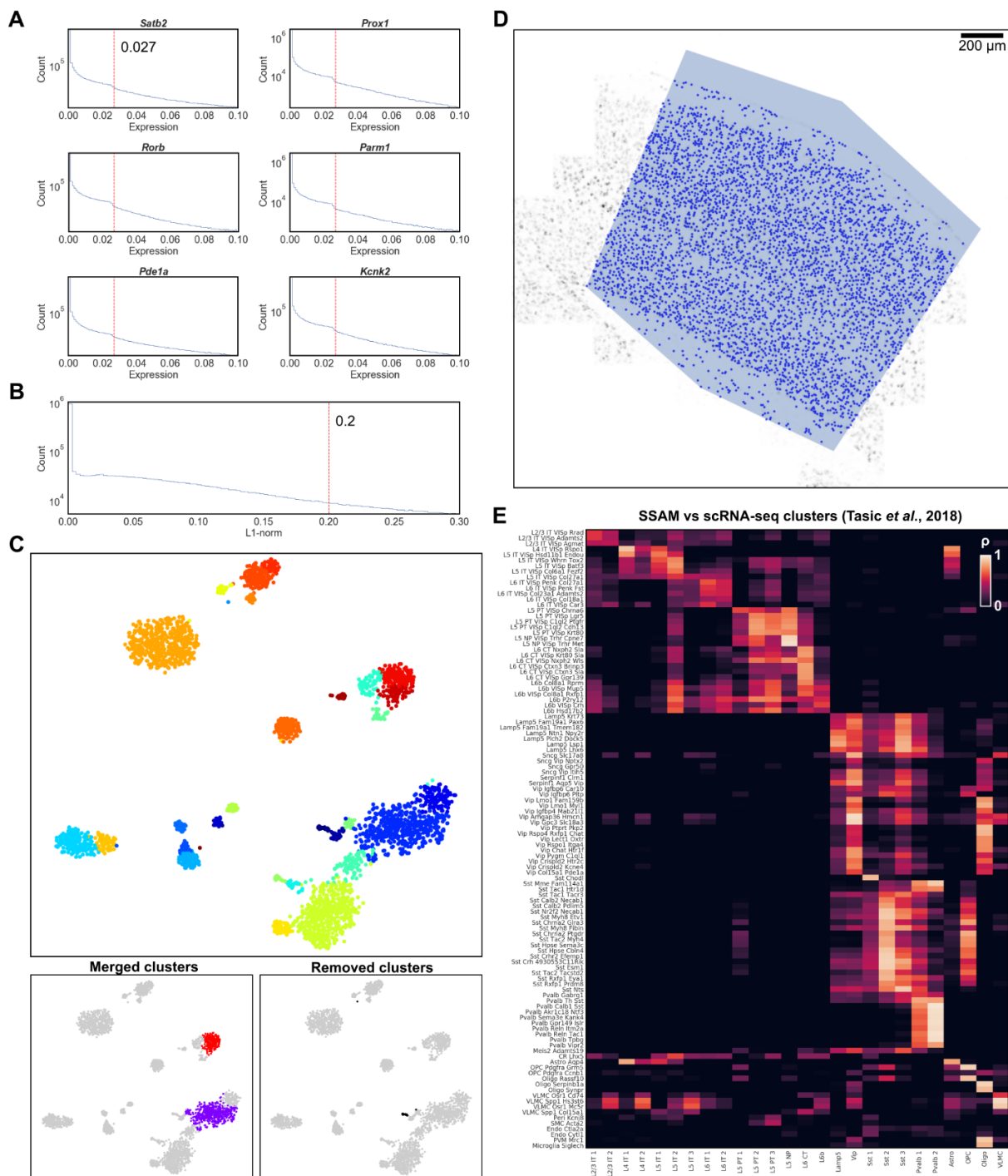


Figure 3.21. Local maxima selection and cell-type signature identification in the mouse *VISp* smFISH dataset.

(A) The gene expression threshold (red vertical line) was defined by the location of an observable drop in the gene expression histogram. Here 6 randomly selected genes are presented; (B) The total gene expression (red vertical line) threshold was defined based on the total gene expression histogram; (C) A 2D t-SNE embedding of clustering local maxima vectors, excluding the ones do not have high correlation to each cluster's medoid (top, see also Figure 3.2B). The

merged clusters and the removed clusters are visualized in the same t-SNE map (bottom); (D) Selected local maxima on the vector field. The VISp region is supplied as in input mask, represented as a polygon on the image so that SSAM only selects local maxima in the VISp region; (E) Comparison of the SSAM *de novo* cell-type signatures to the single-cell RNA sequencing derived cell-type signatures from Tasic et al (2018) [132].

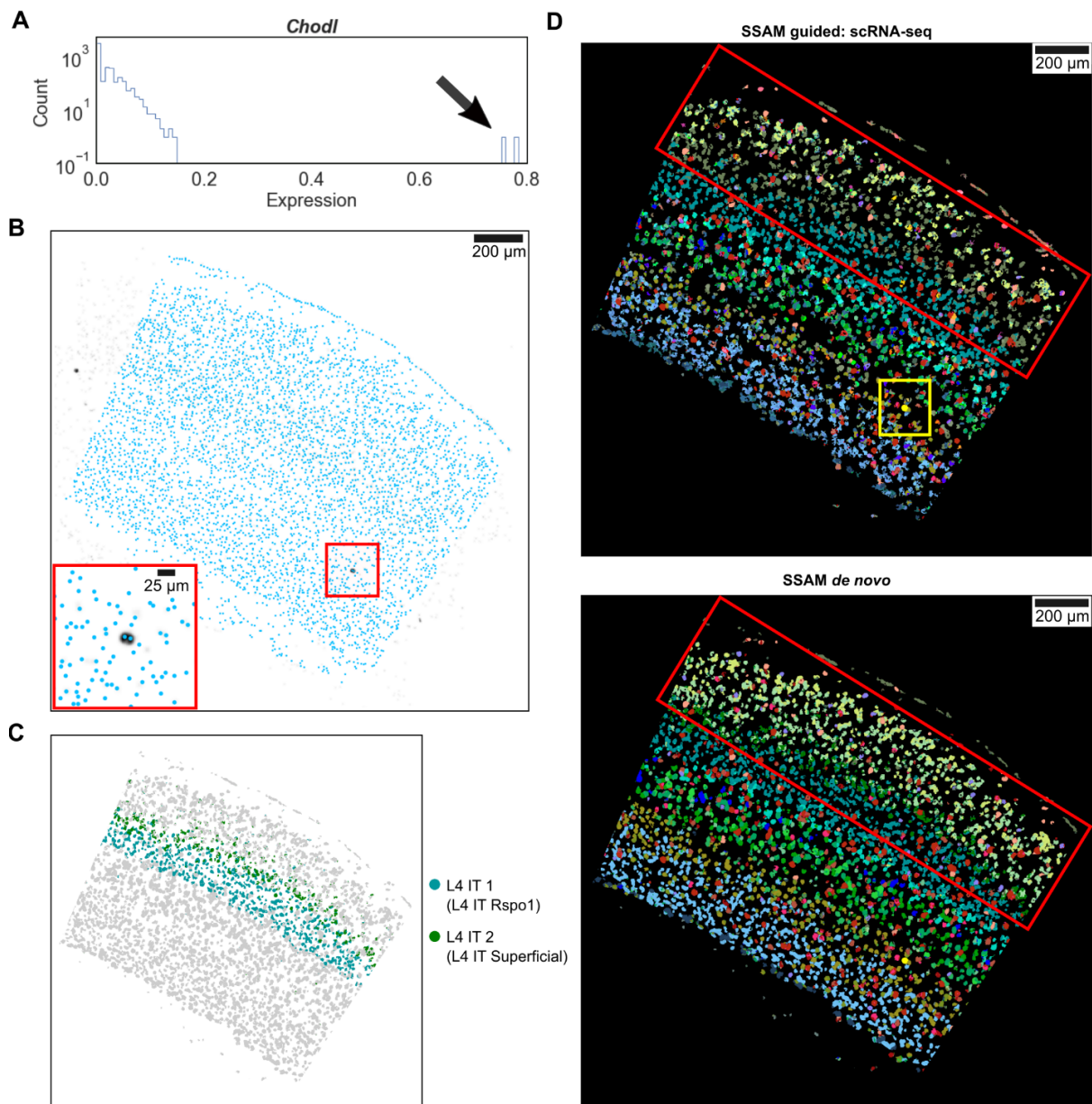


Figure 3.22. Rescuing rare *Sst Chodl* cell types, and identifying new sub-layering in the L4 cortical layer in the mouse VISp smFISH dataset.

(A) Distribution of *Chodl* gene expression in the local maxima vectors. There are two distinctive vectors (marked by arrow), which highly express *Chodl* gene; (B) The location of the two high *Chodl* expressing vectors. The locations of the local maxima vectors (light blue dots) are overlaid on the KDE signal of *Chodl* gene expression. There are 2 dark spots indicating high *Chodl* expression, one inside the selected VISp region (boxed in red), and one outside. Both *Chodl* expressing vectors originate from the same dark spot in the selected tissue region (inset); (C) Localization of the two L4 IT cell types indicates L4 sub-layering. The heterogeneity of L4 IT cell types is related with their localizations in layer 4 of the VISp, with cell type L4 IT1 localizing to the deep portion, and L4 IT2 localizing superficially; (D) Comparison of SSAM

guided mode and *de novo* mode results. The two results are visually similar apart from the L2/3 region (boxed in red), which shows wrongly mapped cell types in the guided mode result, indicating that the guided mode does not always guarantee the best results. Instead, the *de novo* mode can be used to find the accurate mapping of cell types on tissue when guided mode fails to reconstruct cell-type maps correctly. However, SSAM was able to find the rare Sst Chodl cell type in guided mode (boxed in yellow), which resulted in the manual rescue of the Sst Chodl local maxima vectors for cell-type classification in the *de novo* mode.

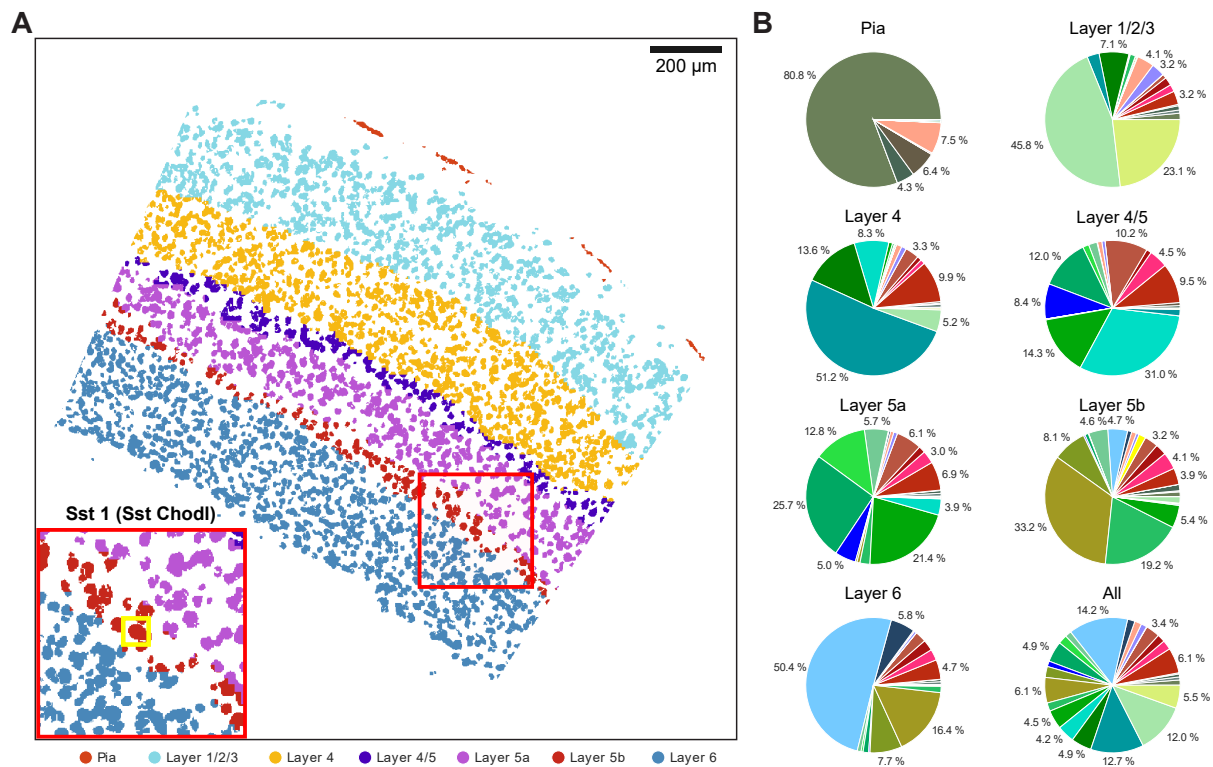


Figure 3.23. Rare Sst Chodl cell type localizes to the L5b cortical layer of the mouse VISp region.

(A) Tissue domain map generated by SSAM. Tissue domain signatures were identified from clustering local cell-type composition over sliding 100 µm circular windows and projected back onto the cell-type map. The reconstruction shows the various cortical layers within the adult mouse VISp, with very clear separation of the pia layer, and separation of layer 5 into 2 layers, 5a and 5b. The inset zooms into the location of the rare Sst Chodl cell type found in layer 5b; (B) Cell-type composition within each tissue domain. The colors in the pie charts correspond to the cell-type legend in Figure 3.20B.

3.2.4 Analysis of mouse visual cortex (VISp) imaged by multiplexed smFISH

Based on the validation of the accuracy and the robustness of SSAM analysis with two publicly available datasets, SSAM was applied to a newly generated data, a mouse brain visual cortex (VISp) imaged by multiplexed smFISH at Allen brain institute, as a part of SpaceTx consortium under the umbrella of Human Cell Atlas project (Figure 3.20, 3.21, 3.22, 3.23). The dataset contains the locations of mRNAs of 22 genes imaged in 2D space (further details can be found in the Methods details section).

Based on the location of mRNAs, KDE was performed to estimate the expression of each gene at the lattice point, resulting in 22 images of the 22 genes. The images were then stacked to form a vector field, and then the representative vectors were selected at the local maxima of the total expression of each vector in the vector field. The vectors were then thresholded with their gene expressions and the total expressions (Figure 3.21A,B). The thresholded vectors were then restricted within the VISp region on the tissue (Figure 3.21D) using a manually defined input mask. The resulting vectors and the whole vector field were normalized with *sctransform* [128].

For *de novo* analysis, the representative vectors were clustered using the Louvain clustering algorithm described in the above sections (Figure 3.20A,B). Based on the diagnostic plots, similar clusters were merged and the clusters mapped to the artifacts were removed (Figure 3.21C). The centroids of the remaining clusters showed a high correlation to that of the clusters from the single-cell RNA sequencing data (Figure 3.21E). The centroids were then mapped to the vector field and merged into the final cell-type map (Figure 3.20C). For guided analysis, the signatures obtained from a single-cell RNA sequencing experiment [132] were mapped to the vector field to form a guided cell-type map.

Despite the high visual similarity between the guided and the *de novo* cell-type maps, there were two major differences (Figure 3.20C). (1) The first difference was the existence of Sst Chodl cell type in tissue. This cell type is known for its rareness among the whole population of cells found by the single-cell RNA sequencing data [133–135], and there is a curiosity of its spatial organization in the tissue. In the resulting cell-type map, the Sst Chodl cell type was found in the cell-type map in the guided mode, whereas it was not able to be found in the *de novo* cell-type map in the first attempt. To find the reason why *de novo* mode could not find this cluster, firstly the representative vectors were reviewed whether it contains the vectors originating from Sst Chodl cell type. Indeed, there were 2 vectors that highly express gene *Chodl*, which is the marker gene for Sst Chodl cell type (Figure 3.22A,B). However, it was revealed that the Louvain clustering algorithm did not identify the two vectors as an independent cluster, even though the difference of the gene expression of the two vectors were clearly different from other vectors. Since it is known that the Louvain community detection algorithm is not

performing well detecting small communities in a network [136], it was concluded that this is a general problem of the Louvain clustering algorithm itself, not a problem of SSAM framework. Still, the Louvain clustering algorithm is the most widely used and provides reasonable results in terms of clustering single-cell RNA sequencing data, alternative clustering algorithm was not considered for this dataset. Instead, the two vectors were simply manually rescued as an individual cluster, and the cell-type map was regenerated based on it (Figure 3.20C). (2) The other difference was the difference in cell types in L2 layer. In the SSAM guided mode, the majority of cell types in L2 layer were mapped to the VLMC type, which makes more sense rather be mapped to a neuronal cell type (Figure 3.22D). This was mainly due to the lack of other neuronal marker genes among the limited number of genes imaged by the experiment, therefore the cluster was mapped to the VLMC cell type which also highly expresses a gene *Alcam*. This was corrected in the *de novo* mode, assigning this cell type to L2 neurons.

Interestingly, there were two different cell-types observed in the L4 layer in the *de novo* cell-type map, which have a similar marker gene expression to each other (Figure 3.20C, 3.22C). Despite the similar gene expression signature, it was clearly observed that one is closer to the superficial layer, and the other one was located below it. However, for both cell types, the closest cell type in the single-cell RNA sequencing data was only one, L4 IT. Instead, it was previously reported that the L4 IT cell type showed high heterogeneity within the tissue [132]. Therefore, it was concluded that there are actually two different L4 IT cell types, but the difference can only be clearly identified in the spatial context, but hard to be distinguished in the previous analysis with the single-cell RNA sequencing data. Here the one closer to the superficial layer was named to L4 IT Superficial in our analysis.

Finally, the tissue domains were found by sweeping the circular window with diameter 100 μm on the *de novo* cell-type map (Figure 3.23). The local composition of cell types was clustered using an agglomerative clustering algorithm to find the tissue domains. The resulting domain map revealed the layered structure of the cortex, which matches well with the previous histological findings.

3.3 Discussion

In this chapter I presented SSAM, a novel computational framework to infer cell types without requiring segmentation. The robustness of SSAM was tested using two publicly available data imaged by osmFISH and MERFISH, and one newly generated data, multiplexed smFISH.

Importantly, the cell-type map generated by SSAM with the two publicly available data showed not only the similarity to the previously reported results but also clear improvements in terms of finding the spatial organization of cell types. Especially with MERFISH data, it was shown that SSAM can be easily extended to 3D space to generate a 3D cell-type map,

which can be used to investigate the organization of cell types in 3D. When SSAM was applied to the multiplexed FISH data, SSAM was able to unravel the previously unknown source of heterogeneity of the L4 IT cell type, which led to a discovery of a new cell type that was able to be clearly distinguished in the spatial context.

In the osmFISH data, we found one interesting Astrocyte cell type as a result of the *de novo* analysis. The cell type highly expresses the gene *Aldoc* but not so much of other marker genes of astrocytes, like *Mfge8* or *Gfap*. The cluster was not able to be merged with other astrocyte clusters, which only weakens the correlation between the vector field and the cluster centroid. Therefore, the cluster was named as Astrocyte Aldoc. However, morphologically the resulting Astrocyte Aldoc seemed to be a part of Astrocyte Mfge8. Therefore, it was assumed that the Astrocyte Aldoc could be a part of Astrocyte Mfge8 instead of being a new cell type, due to an unknown subcellular internal organization of cells. Since it has been reported that such organizations can be observed using imaging methods [137], SSAM has the potential to discover them by comparing the cell-type map with the previous findings to uncover the mystery of subcellular organizations, or possibly further post-transcriptional regulations associated with them [138].

3.3.1 KDE bandwidth and lattice spacing

One possible concern of SSAM is finding the most appropriate bandwidth of KDE and the lattice spacing to estimate spatial gene expressions on the lattice points. Here the bandwidth determines the width of the Gaussian kernel, therefore it determines the amount of ‘smoothness’ after the estimation, and the lattice spacing determines the resolution of the final cell-type map (Figures 3.24-3.27). Therefore the lattice spacing does not play a critical role, but the bandwidth does – for example, too high bandwidth makes the signal too much smoothed, which can interfere the gene expression signature of the nearby cells, whereas too low bandwidth makes harder to reconstruct the shape of cells in the resulting cell-type map. For the three datasets described in this chapter, the bandwidth was set to 2.5 μm which makes the full width tenth maximum (FWTM) of Gaussian kernel to be similar to 10 μm , under the assumption that the average cell diameter is around 10 μm for the three examples. With this value, the signal interference was not a big problem to infer the cell types. Furthermore, the robustness of SSAM *de novo* cell-type calling procedure was tested with the different bandwidths except 2.5 μm (0.5, 1, 5, 10 μm , Figure 3.28). Even with extreme cases like 0.5 or 10 μm of bandwidth, the resulting clusters still preserved high correlation to the *de novo* clusters found with bandwidth 2.5 μm , which means that the same number of clusters could be still found with different bandwidths after a proper merging step (except 10 μm , which lost 2 clusters) – this confirms that the bandwidth does not play a big role in terms of the cell-type calling. The only remaining problem is

Figures 3.24-3.27. Effect of KDE bandwidth and lattice spacing in 4 different regions.

The figures show the cell-type map of 4 different regions (highlighted in white boxes), generated with 5 different KDE bandwidth (0.5, 1, 2.5, 5, 10 μm) and 3 different lattice spacing (0.5, 1, 2.5 μm) in guided-mode SSAM (guided by segmentation-based signatures) using osmFISH dataset. The figure shows that only the bandwidth is related to the size of detected blobs cell-type map, and the lattice spacing is only related to its pixel size. For all cases, even including very extreme cases with bandwidths 0.5 and 10 μm , the pixels in the cell-type map correlates well with the corresponding cell-type signatures and colored correctly. In other words, the signals estimated by KDE preserves most of the cell-type signatures even with extreme change of bandwidth, which implies that the cell-type identification will not be largely affected by a small change of bandwidth near 2.5 μm .

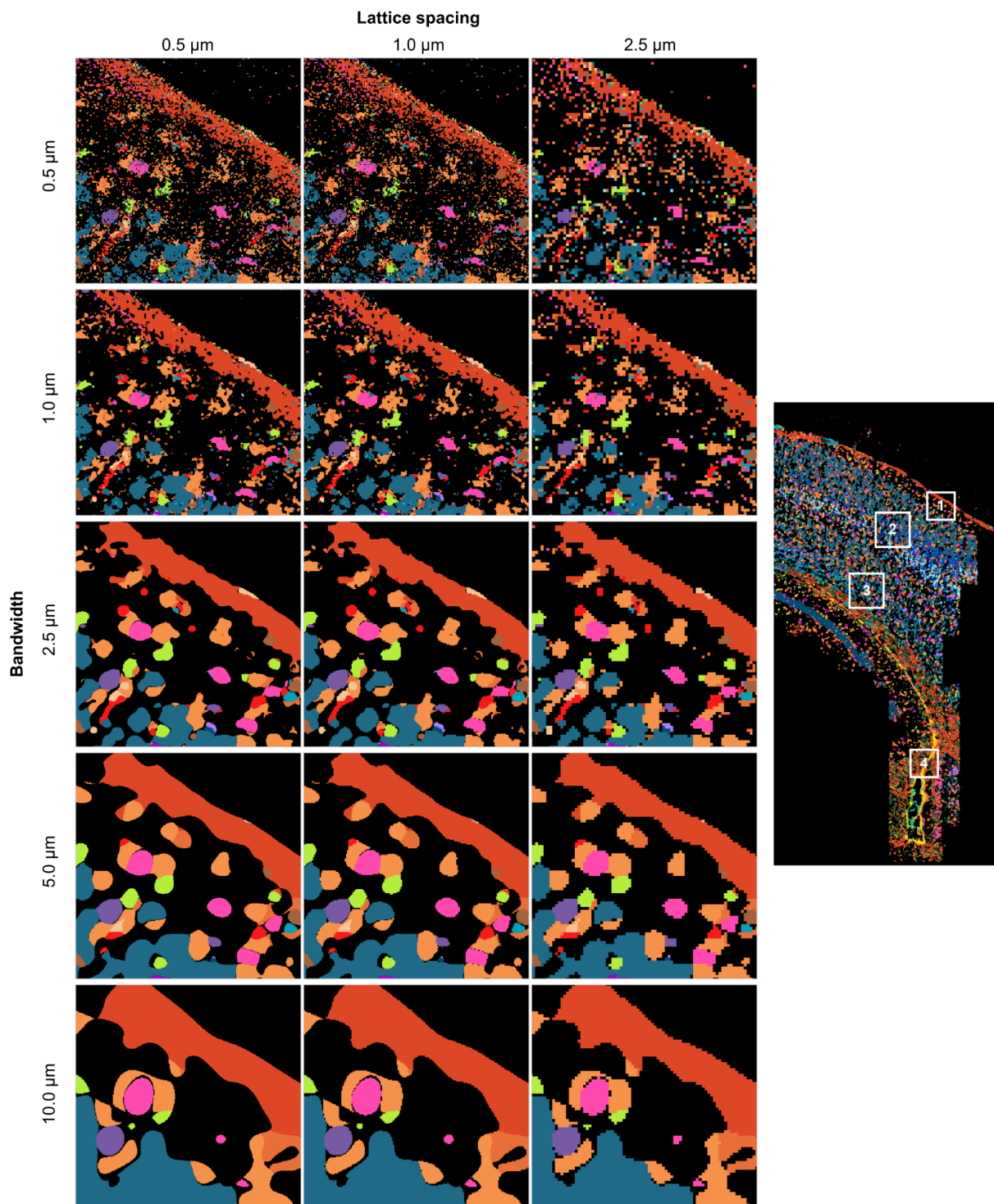


Figure 3.24. Effect of KDE bandwidth and lattice spacing in the first region.

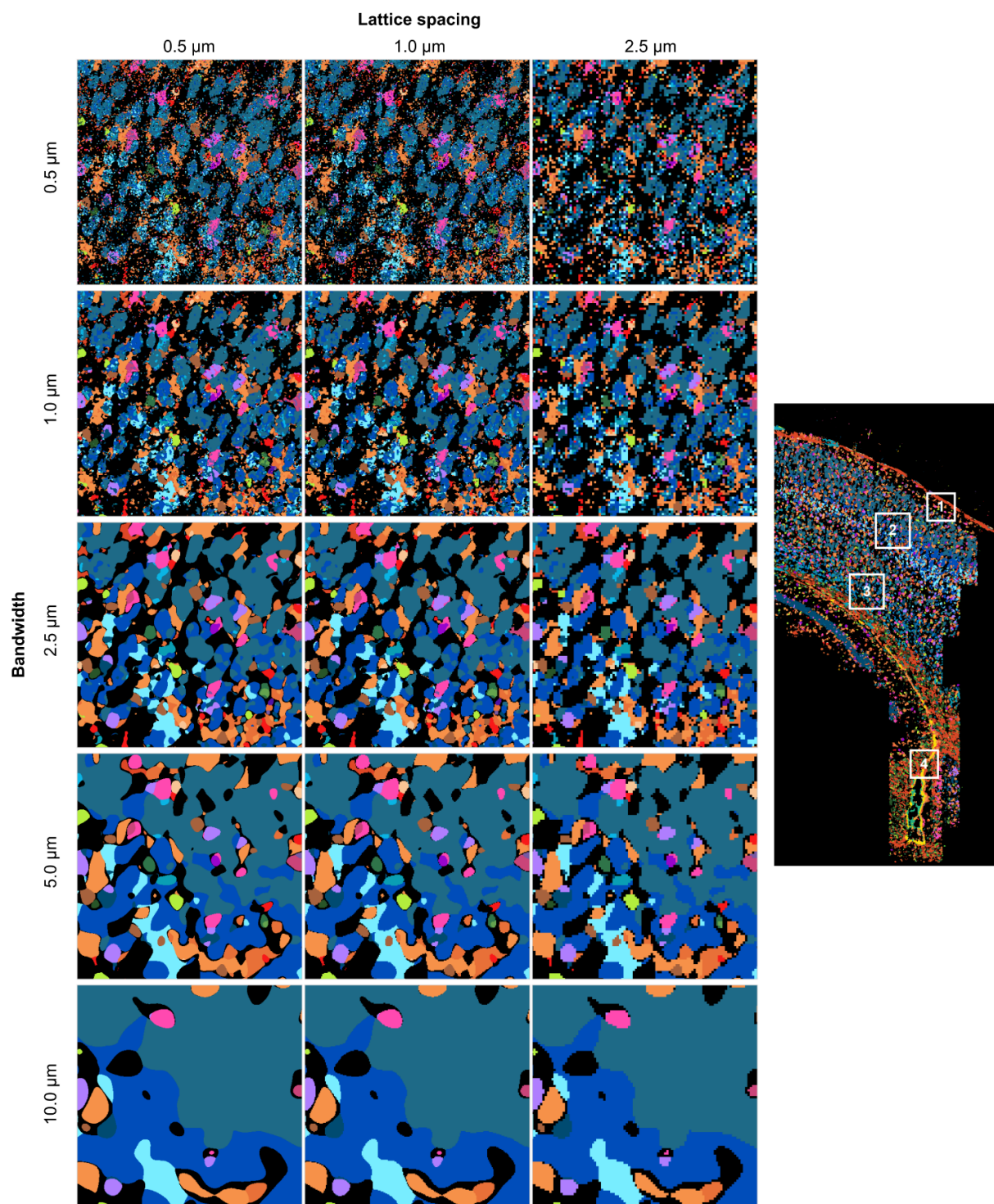


Figure 3.25. Effect of KDE bandwidth and lattice spacing in the second region.

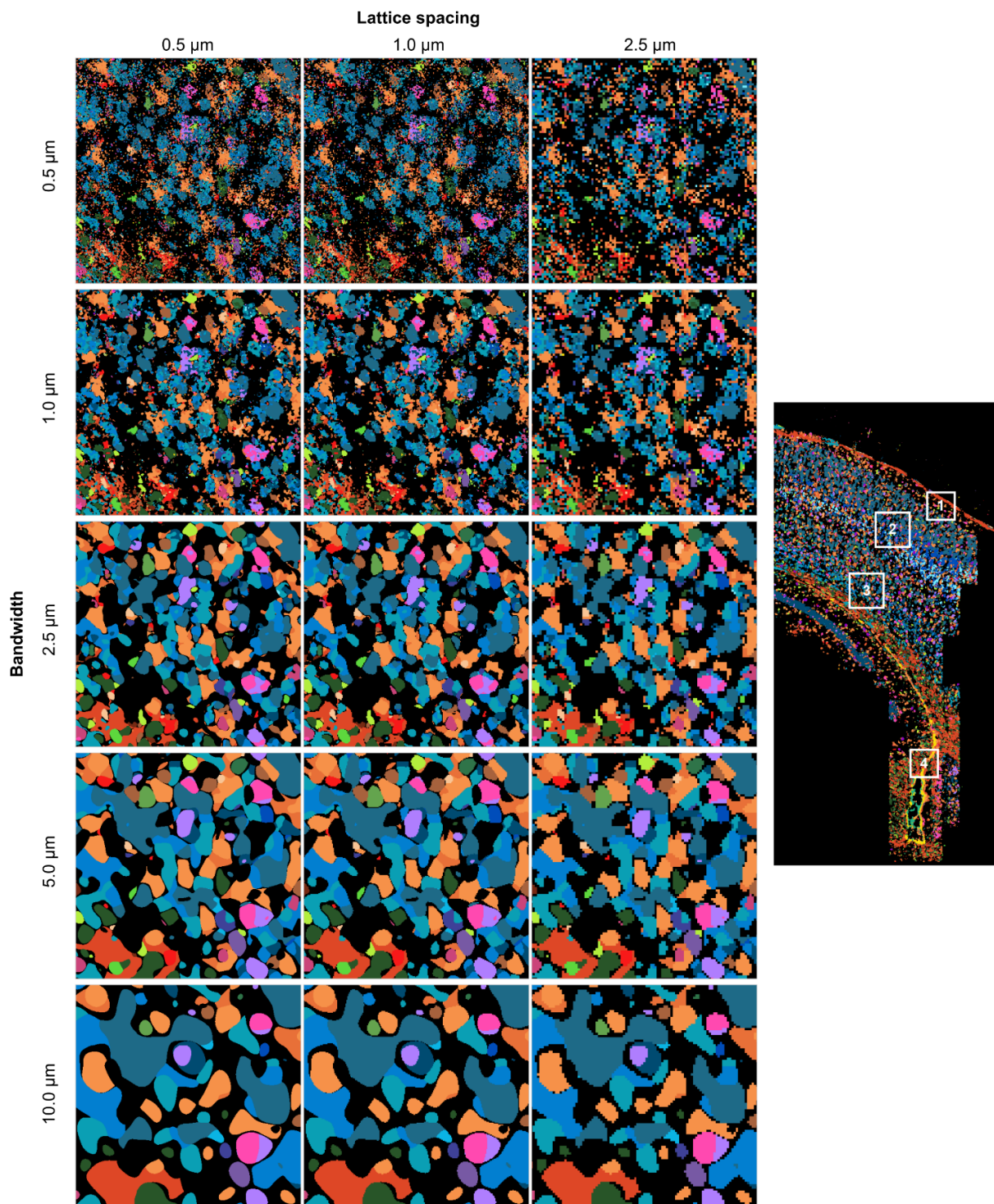


Figure 3.26. Effect of KDE bandwidth and lattice spacing in the third region.

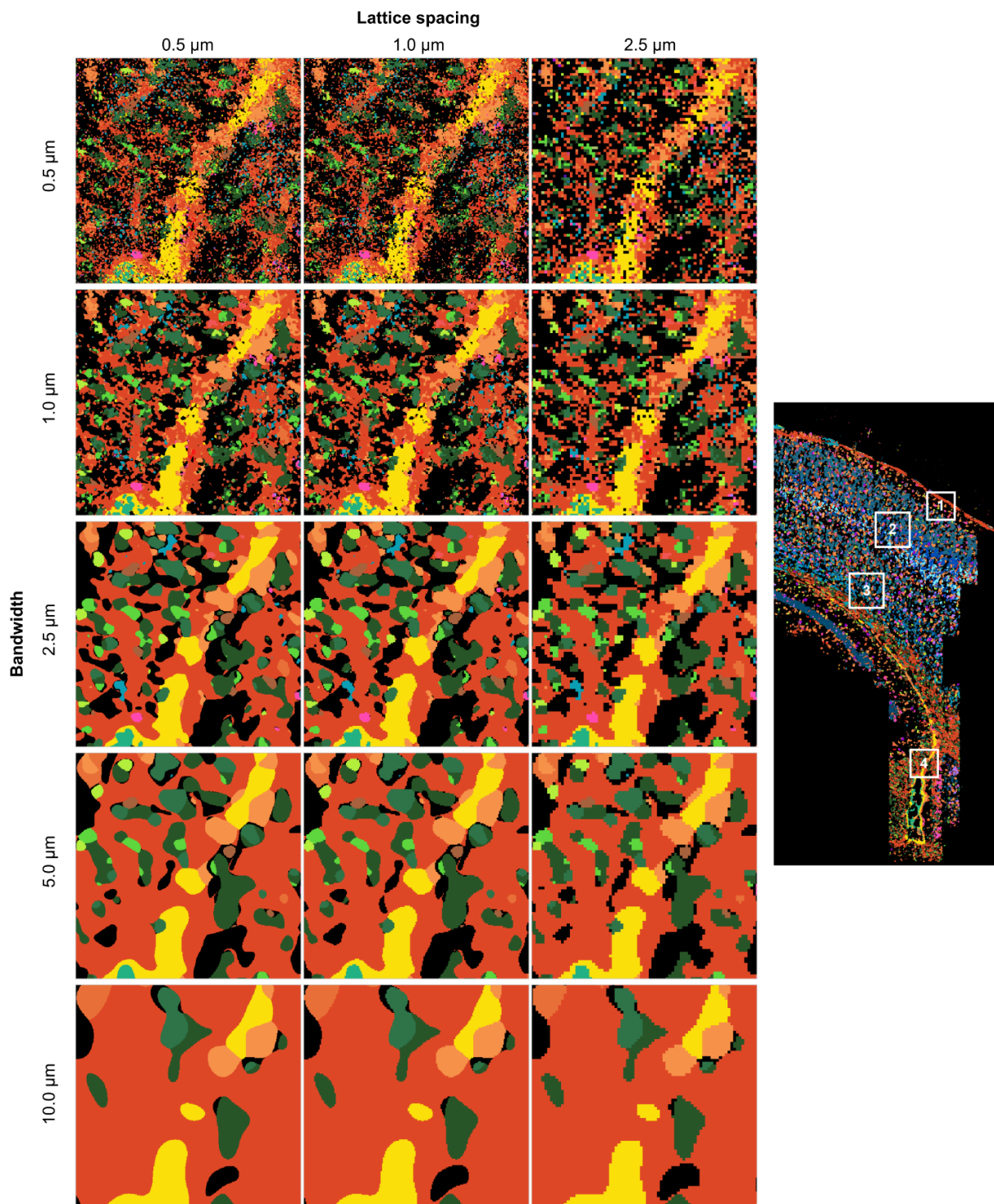


Figure 3.27. Effect of KDE bandwidth and lattice spacing in the fourth region.

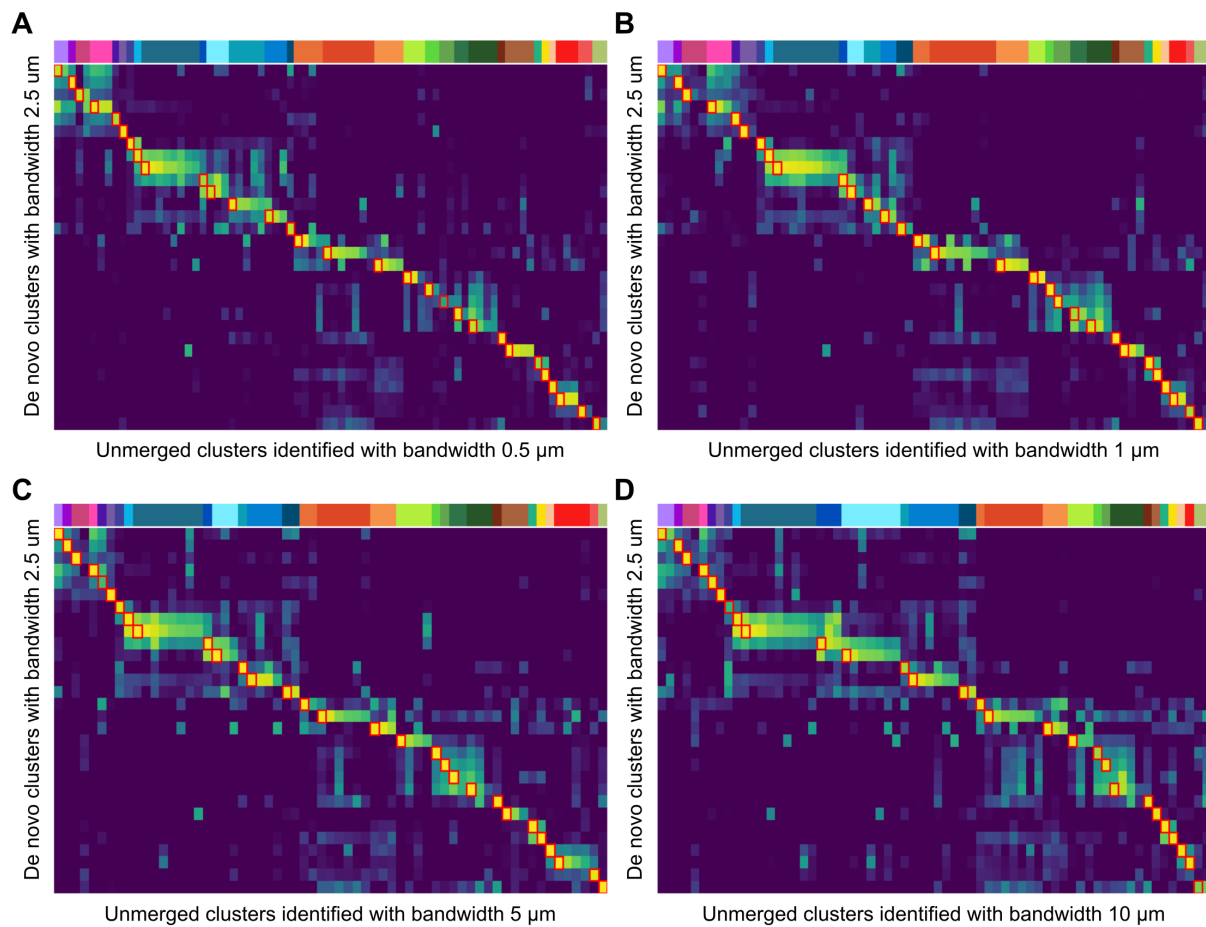


Figure 3.28. Correlations between the *de novo* clusters with bandwidth 2.5 and the unmerged clusters with different bandwidths.

(All panels) Pearson's correlation between the clusters with bandwidth 2.5 and the unmerged clusters identified with different bandwidths, 0.5, 1, 5, and 10. The representative vectors were selected at the same locations identified in the original *de novo* analysis. The highest correlated clusters for each cluster identified in the original *de novo* analysis are marked with red boxes; (A) Among the unmerged 76 clusters with bandwidth 0.5, all the *de novo* clusters were highest correlated with at least one of them. The highest correlated values in red boxes were with max 0.997, min 0.607, and median 0.961; (B) Among the unmerged 67 clusters with bandwidth 1, all the *de novo* clusters were highest correlated with at least one of them. The highest correlated values in red boxes were with max 0.998, min 0.861, and median 0.990; (C) Among the unmerged 63 clusters with bandwidth 5, all the *de novo* clusters were highest correlated with at least one of them. The highest correlated values in red boxes were with max 0.998, min 0.690, and median 0.993; (D) Among the unmerged 66 clusters with bandwidth 10, except 2 clusters, all the other *de novo* clusters were highest correlated with at least one of them. The highest correlated values in red boxes were with max 0.993, min 0.658, and median 0.968.

that too large bandwidth makes the blobs very large in the resulting cell-type map. Based on the assumption that the average diameter of the cells will be 10 μm for the many tissues, therefore the default bandwidth of SSAM was set to be 2.5 μm . Thus, it is strongly encouraged that the users test different bandwidths if the average cell size is significantly different from 10 μm to generate a better cell-type map.

3.3.2 Possible extension of SSAM for *in situ* sequencing methods

In this chapter, only the multiplexed FISH datasets with up to tens or hundreds of genes per experiment were discussed, but recent advances on the techniques dramatically increased the number of genes which can be imaged, up to the order of tens of thousands. Therefore, it is not a dream to replace such techniques to single-cell RNA sequencing to reveal the spatial heterogeneity of the single cells in a tissue. Moreover, there are several techniques developed to sequence mRNAs *in situ*, theoretically, an unlimited number of genes can be imaged per experiment, although shows a lower sensitivity of capturing mRNAs than FISH methods. Despite the low sensitivity, some preliminary SSAM analysis showed that SSAM also works with such *in situ* sequencing [139] (by Mr. Sebastian Tiesmeyer) and STARmap data [119] (Figure 3.29, successfully reconstructing cell-type map based on the datasets.

SSAM is mainly written in Python, and some core functions are written in C to accelerate the computation speed. All SSAM functions are accessible as a Python module, and usage of SSAM is available as Jupyter notebook so that one can follow the analysis steps easily. The source code of SSAM and the example Jupyter notebooks are available on Github (<https://github.com/eislabs/ssam> and https://github.com/eislabs/ssam_example).

In summary, I present SSAM which is a segmentation-free method to infer cell types, and further determine spatial organizations of the inferred cell types in the spatial context via a cell-type map. SSAM can be used to quickly determine the spatial location of certain cell types found by foreign analysis via guided mode, but also can be used to identify new cell types in *de novo* mode. SSAM is planned to be integrated into the Starfish pipeline, which is being developed as a suggested standard processing pipeline to process spatial data in the SpaceTx consortium under the umbrella of the Human Cell Atlas project.

3.3.3 Method details

Using Kernel Density Estimation to generate the gene expression vector field

We used the n-dimensional KDE algorithm to estimate the density of mRNAs in 2D and 3D. To compute Gaussian KDE, we used our own implementation of the KDE algorithm for rapid computation. Spatial distribution of the probability of mRNA presence is estimated using the kernel density estimation;

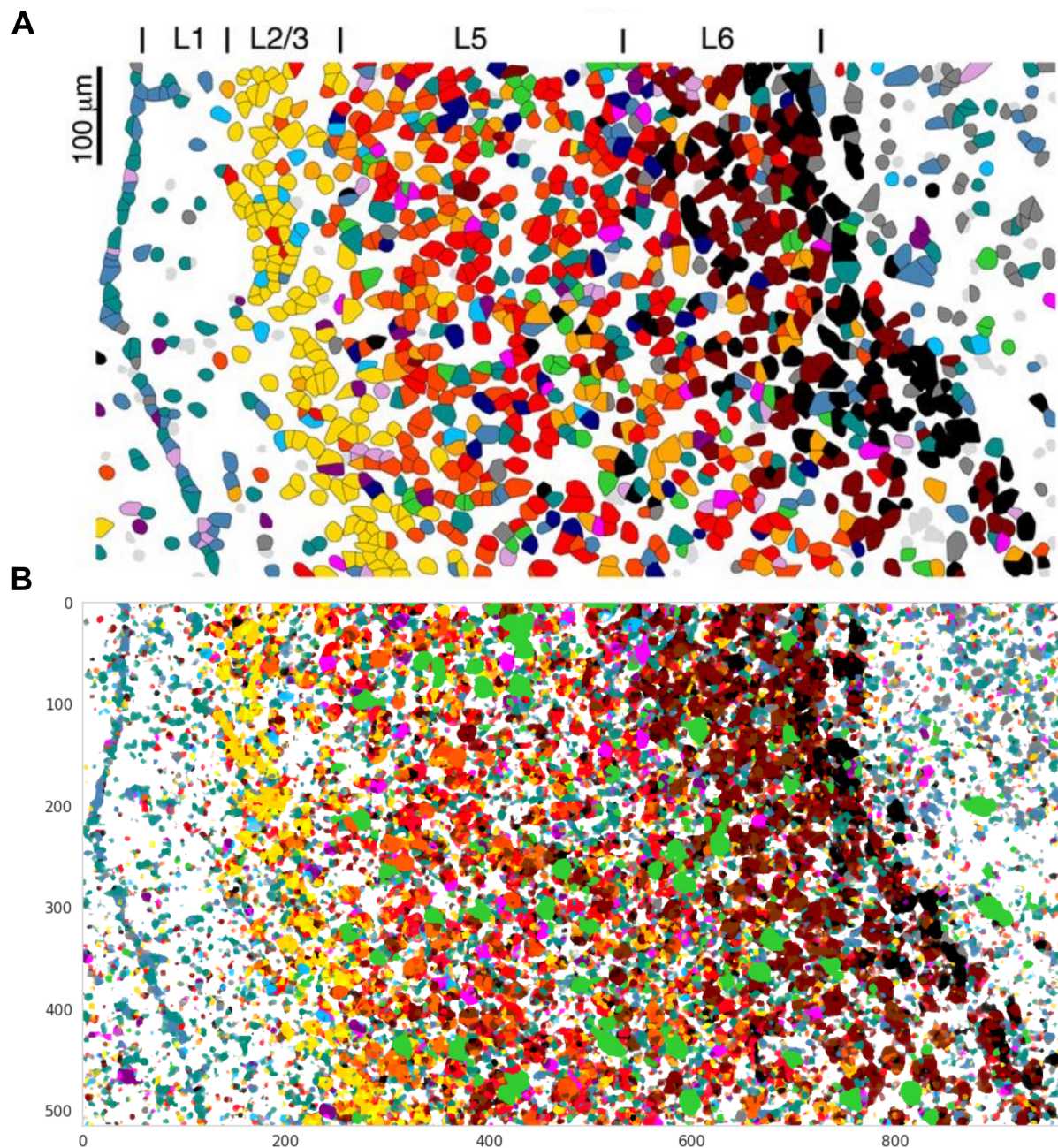


Figure 3.29. The preliminary cell-type map with mPFC imaged by STARmap.

(A) This subfigure was taken from the publication of STARmap [119], Figure 4C. This shows the previously published spatial organization of the cell types in mPFC region; (B) SSAM guided mode cell-type map, guided by the signatures determined in Wang *et al* based on the segmentation algorithm. The resultant cell-type map shows high visual similarity to the one that was previously published.

$$\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \kappa_h(\mathbf{x} - \mathbf{x}_i)$$

where:

κ_h : kernel function

h : bandwidth

N : number of mRNAs of the same gene

And here we use the Gaussian kernel:

$$\kappa_h(\mathbf{x}) = \frac{1}{(2\pi h^2)^{d/2}} e^{-\|\mathbf{x}\|^2/2h^2}$$

where:

d : dimension

Note that each data point \mathbf{x} lies within the respective image, hence the dimension d is either two or three. Ideally, the probability density at each lattice point must be evaluated by integration over the unit area. The lattice size is considered sufficiently fine-grained to capture all relevant information on the continuous Gaussian curve. To create a proper probability density, the lattice points are scaled to a sum of 1. Finally, the gene expression is estimated by multiplying each density by its total number of mRNA molecules.

Normalization of local maxima vectors and the vector field

Since the gene expression profiles of local maxima vectors are representative of the transcriptomes of cells, we considered them to be analogous to the gene expression count matrix obtained from single-cell RNA sequencing (scRNA-seq) using unique molecular identifiers (UMI). Therefore, we normalized the local maxima vectors of the vector field (which would be representative of single cells) using *sctransform* [128], a normalization and regularization algorithm for UMI count data. After that, each vector of the vector field is normalized using *sctransform*, with the same parameters previously used to normalize the local maxima.

Clustering of representative gene expression vectors

The clustering algorithm implemented in SSAM is based on the source code of the R package Seurat [86]. Here, we used the same algorithm reimplemented in Python. In short, an SNN network with a correlation metric is built using a python package NetworkX [140]. The weight of the network is calculated by a Jaccard similarity coefficient. A weight smaller than 1/15

was set to zero. Clustering was done by detecting communities in the network using a Louvain community detection algorithm implemented in Python (python-louvain, <https://python-louvain.readthedocs.io/>). The resolution of the Louvain algorithm is set to 0.15.

SSAM diagnostic plots

To provide support to the user on whether to merge or remove clusters, SSAM generates a cluster-wise ‘diagnostic plot’, which consists of the following four panels: 1) location of the vectors originating from the cluster, 2) a map of the centroid embedded into the vector field, 3) the centroid of the cluster, 4) the location of the cluster in t-SNE or UMAP embedding. In the three applications in this chapter, the clusters to be merged or removed often showed a mismatch between the location of vectors (panel 1) and the map of the centroid (panel 2). For sub-cell types, the map typically does not clearly show the full shape of the cells but only fragments but simultaneously having clear marker gene expression (panel 3). This usually indicates that there is another centroid that has a higher correlation to the expression profile of the entire cell body. In such cases, such centroids are merged to the centroid with a higher correlation. For dubious clusters, it is observed that vectors are usually located outside the tissue region or represent image artifacts (panel 1), the map clearly shows that the centroid is mapped to the artifacts (panel 2), or that the gene expression does not show any clear expression of marker genes (panel 3). Such clusters are removed thereafter. The remaining clusters are then identified by comparing cluster marker genes to known cell-type markers. Note that in many cases, the identity of clusters can be easily assigned by comparing the centroids of the clusters to the known cell-type signatures, e.g., from single-cell RNA sequencing. Therefore, if such signatures are given, SSAM additionally shows the closest cell-type signature among the given signature in the diagnostic plot to help users easily assign classes to clusters. The diagnostic plots for osmFISH, MERFISH, and multiplexed smFISH data are available online in the Jupyter notebook uploaded to Zenodo (<http://doi.org/10.5281/zenodo.3478502>).

Quantification of mRNA abundance in astrocytes and other brain cell types for osmFISH data interpretation

(Note: This work was done by Mr. Sebastian Tiesmeyer, and this subsection was originally written by him.)

The “L5_All.loom” loom object containing scRNA-seq expression data of half a million cells from the mouse nervous system [130] was downloaded. Using Python, the total mRNA molecules per cell were extracted and aggregated by their level 2 class labels (astrocytes, immune, vascular, ependymal, neuronal, peripheral glia and oligodendrocyte cells). The total mRNA counts per class were log-normalized and subsequently followed a normal distribution

(tested using the Shapiro-Wilk test for normality, all p-values $< 1 \times 10^{-4}$ for each class), therefore a Student's t-test was applicable. For each of the two classes of interest ('Astrocytes', 'Immune'), we performed independent log-space t-tests for unequal sample sizes and unequal variance against each of the other classes. Both astrocyte and immune cell classes have significantly lower mRNA molecule counts compared to other cell types (all p-values $< 1 \times 10^{-12}$). While the distribution of mRNA counts in log space followed a normal distribution, the use of a Student's t-test for large numbers may not be appropriate. Hence, we also describe the difference in their distributions. For both astrocyte and immune cell classes, more than half of the cells of each class exhibited a lower UMI count than the lowest quartile of any other cell class.

3D modeling of MERFISH cell-type maps

Firstly, the connected components in 3D were determined using a python package 'connected-components-3d' (<https://github.com/seung-lab/connected-components-3d>). Components comprising fewer than 100 voxels were removed. After this, the voxels filling connected components were removed, and only the contours were used for the vertex of the 3D models. For each vertex the vertex normal was calculated by simple physics simulation, assuming that the direction of vertex normal vector is the same as the force vector when there are pushing forces between all of the contour voxels. The surface of the objects is reconstructed using a screened Poisson reconstruction algorithm [141, 142] using default parameters. The number of vertices was reduced to 5 % of the total number of vertices using 'vtkQuadricDecimation' function [143, 144] of VTK library [145]. Finally, the objects are merged into one file. Each scene of the rotating movie was created using Meshlab [146].

VISP multiplexed smFISH data generation

(Note: This subsection was originally written by the data provider, the team at Allen Brain Institute.)

Multiplexed smFISH data of mouse primary visual cortex (VISp) was generated as part of the SpaceTx consortium. Tissue processing was carried out as previously described [147], with some modifications.

Silanization of coverslips (#1.5, Thorlabs CG15KH) was performed by plasma cleaning for 30 min in a Plasma-Prep III (SPI 11050-AB), followed by vapor deposition of 3-aminopropyltriethoxysilane (APES, Sigma A3648) in a vacuum for 10 minutes. Coverslips were then washed in 100 % methanol for 2 x 5 minutes, allowed to dry, and stored in a dust-free environment until use. s Fresh-frozen mouse brain tissue was sectioned at 10 μm onto silanized coverslips, let dry for 20 min at -20°C , then fixed for 15 min at 4°C in 4 % PFA in PBS. Sections were

washed 3×10 min in PBS, then permeabilized and dehydrated with chilled 100 % methanol at -20 °C for 10 min and allowed to dry. Sections were stored at -80 °C until use. Frozen sections were rehydrated in 2X SSC (Sigma 20XSSC, 15557036) for 5 min, then treated 10 min with 8 % SDS (Sigma 724255) in PBS at room temperature. Sections were washed 5 times in 2X SSC. Sections were then incubated in hybridization buffer (10 % Formamide (v/v, Sigma 4650), 10 % dextran sulfate (w/v, Sigma D8906), 200 μ g/mL BSA (ThermoFisher AM2616), 2 mM ribonucleoside vanadyl complex (New England Biolabs S1402S), 1 mg/ml tRNA (Sigma 10109541001) in 2X SSC) for 5 min at 37°C. Probes were diluted in hybridization buffer at a concentration of 250 nM and hybridized at 37°C for 2 h. Following hybridization, sections were washed 2×10 min at 37°C in wash buffer (2X SSC, 20 % Formamide), and 1×10 min in wash buffer with 5 μ g/ml DAPI (Sigma 32670), then washed 3 times with 2X SSC. Sections were then imaged in Imaging buffer (20 mM Tris-HCl pH 8, 50 mM NaCl, 0.8 % glucose (Sigma G8270), 30 U/ml pyranose oxidase (Sigma P4234), 50 μ g/ml catalase (Abcam ab219092)). Following imaging, sections were incubated 3×10 min in stripping buffer (65 % formamide, 2X SSC) at 30°C to remove hybridization probes from the first round. Sections were then washed in 2X SSC for 3×5 min at room temperature before repeating the hybridization procedure.

The multiplexed smFISH image data was collected and processed using methods previously described [147], except that images from different rounds of hybridization were registered in (x,y) based on the DAPI signal. The spot locations and raw data are available on request.

Plotting

The python packages Matplotlib 3.1.0 [148] and Seaborn 0.9.0 [149] were used to draw 2D images, plots, and heatmaps. In SSAM, helper functions are included to easily generate plots.

Movies

Movies were generated by using Virtualdub (1.10.4-AMD64, <http://www.virtualdub.org/>). The H.264 codec was used to compress videos.

Software

Python version 3.7.0 was used throughout. The following python packages were used: numpy, scipy, pandas, matplotlib, seaborn, scikit-learn, umap-learn, python-louvain, sparse, scikit-image. R package sctransform was used for normalization and variance stabilization of the data.

3.4 Contributed publications

- Jeongbin Park*, Wonyl Choi*, Sebastian Tiesmeyer, Brian Long, Lars E Borm, Emma Garren, Thuc Nghi Nguyen, Simone Codeluppi, Matthias Schlesner, Bosiljka Tasic, Roland Eils, and Naveed Ishaque. Segmentation-free inference of cell types from *in situ* transcriptomics data. Under review.
 - Accepted as an official Human Cell Atlas publication.
 - Preprint available at BioRxiv, <https://doi.org/10.1101/800748>

*: Shared first authors.

Bibliography

- [1] Luciano A Marraffini and Erik J Sontheimer. Self versus non-self discrimination during crispr rna-directed immunity. *Nature*, 463(7280):568, 2010.
- [2] Francisco J.M. Mojica, Chcsar Díez-Villaseñor, Jesús García-Martínez, and Elena Soria. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution*, 60(2):174–182, Feb 2005.
- [3] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821, 2012.
- [4] Giedrius Gasiunas, Rodolphe Barrangou, Philippe Horvath, and Virginijus Siksnys. Cas9–crRNA ribonucleoprotein complex mediates specific dna cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences*, 109(39):15539–15540, 2012.
- [5] Seung Woo Cho, Sojung Kim, Jong Min Kim, and Jin-Soo Kim. Targeted genome engineering in human cells with the cas9 rna-guided endonuclease. *Nature biotechnology*, 31(3):230, 2013.
- [6] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano A Marraffini, et al. Multiplex genome engineering using crispr/cas systems. *Science*, 339(6121):819–823, 2013.
- [7] Prashant Mali, Luhan Yang, Kevin M Esvelt, John Aach, Marc Guell, James E DiCarlo, Julie E Norville, and George M Church. Rna-guided human genome engineering via cas9. *Science*, 339(6121):823–826, 2013.
- [8] Woong Y Hwang, Yanfang Fu, Deepak Reyon, Morgan L Maeder, Shengdar Q Tsai, Jeffrey D Sander, Randall T Peterson, JR Joanna Yeh, and J Keith Joung. Efficient genome editing in zebrafish using a crispr-cas system. *Nature biotechnology*, 31(3):227, 2013.

- [9] Wenyan Jiang, David Bikard, David Cox, Feng Zhang, and Luciano A Marraffini. Rna-guided editing of bacterial genomes using crispr-cas systems. *Nature biotechnology*, 31(3):233, 2013.
- [10] Martin Jinek, Alexandra East, Aaron Cheng, Steven Lin, Enbo Ma, and Jennifer Doudna. Rna-programmed genome editing in human cells. *elife*, 2:e00471, 2013.
- [11] Maximilian Müller, Ciaran M Lee, Giedrius Gasiunas, Timothy H Davis, Thomas J Cradick, Virginijus Siksnys, Gang Bao, Toni Cathomen, and Claudio Mussolino. *Streptococcus thermophilus* crispr-cas9 systems enable specific editing of the human genome. *Molecular Therapy*, 24(3):636–644, 2016.
- [12] Zhonggang Hou, Yan Zhang, Nicholas E Propson, Sara E Howden, Li-Fang Chu, Erik J Sontheimer, and James A Thomson. Efficient genome engineering in human pluripotent stem cells using cas9 from *neisseria meningitidis*. *Proceedings of the National Academy of Sciences*, 110(39):15644–15649, 2013.
- [13] F Ann Ran, Le Cong, Winston X Yan, David A Scott, Jonathan S Gootenberg, Andrea J Kriz, Bernd Zetsche, Ophir Shalem, Xuebing Wu, Kira S Makarova, et al. In vivo genome editing using *staphylococcus aureus* cas9. *Nature*, 520(7546):186, 2015.
- [14] Eunji Kim, Taeyoung Koo, Sung Wook Park, Daesik Kim, Kyoungmi Kim, Hee-Yeon Cho, Dong Woo Song, Kyu Jun Lee, Min Hee Jung, Seokjoong Kim, et al. In vivo genome editing with a small cas9 orthologue derived from *campylobacter jejuni*. *Nature communications*, 8:14500, 2017.
- [15] Alireza Edraki, Aamir Mir, Raed Ibraheim, Ildar Gainetdinov, Yeonsoo Yoon, Chun-Qing Song, Yueying Cao, Judith Gallant, Wen Xue, Jaime A Rivera-Pérez, et al. A compact, high-accuracy cas9 with a dinucleotide pam for in vivo genome editing. *Molecular cell*, 73(4):714–726, 2019.
- [16] Benjamin P Kleinstiver, Michelle S Prew, Shengdar Q Tsai, Ved V Topkar, Nhu T Nguyen, Zongli Zheng, Andrew PW Gonzales, Zhuyun Li, Randall T Peterson, Jing-Ruey Joanna Yeh, et al. Engineered crispr-cas9 nucleases with altered pam specificities. *Nature*, 523(7561):481, 2015.
- [17] Johnny H Hu, Shannon M Miller, Maarten H Geurts, Weixin Tang, Liwei Chen, Ning Sun, Christina M Zeina, Xue Gao, Holly A Rees, Zhi Lin, et al. Evolved cas9 variants with broad pam compatibility and high dna specificity. *Nature*, 556(7699):57, 2018.

- [18] Jonathan Strecker, Sara Jones, Balwina Koopal, Jonathan Schmid-Burgk, Bernd Zetsche, Linyi Gao, Kira S Makarova, Eugene V Koonin, and Feng Zhang. Engineering of crispr-cas12b for human genome editing. *Nature communications*, 10(1):212, 2019.
- [19] Bernd Zetsche, Jonathan S Gootenberg, Omar O Abudayyeh, Ian M Slaymaker, Kira S Makarova, Patrick Essletzbichler, Sara E Volz, Julia Joung, John Van Der Oost, Aviv Regev, et al. Cpf1 is a single rna-guided endonuclease of a class 2 crispr-cas system. *Cell*, 163(3):759–771, 2015.
- [20] Michal A. Świat, Sofia Dashko, Maxime den Ridder, Melanie Wijsman, John van der Oost, Jean-Marc Daran, and Pascale Daran-Lapujade. FnCpf1: a novel and efficient genome editing tool for *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 45(21):12585–12598, 11 2017.
- [21] Linyi Gao, David BT Cox, Winston X Yan, John C Manteiga, Martin W Schneider, Takashi Yamano, Hiroshi Nishimasu, Osamu Nureki, Nicola Crosetto, and Feng Zhang. Engineered cpf1 variants with altered pam specificities. *Nature biotechnology*, 35(8):789, 2017.
- [22] Hyongbum Kim and Jin-Soo Kim. A guide to genome engineering with programmable nucleases. *Nature Reviews Genetics*, 15(5):321–334, 2014.
- [23] Sangsu Bae, Jiyeon Kweon, Heon Seok Kim, and Jin-Soo Kim. Microhomology-based choice of cas9 nuclease target sites. *Nature methods*, 11(7):705–706, 2014.
- [24] Grégoire Cullot, Julian Boutin, Jérôme Toutain, Florence Prat, Perrine Pennamen, Caroline Rooryck, Martin Teichmann, Emilie Rousseau, Isabelle Lamrissi-Garcia, Véronique Guyonnet-Duperat, et al. Crispr-cas9 genome editing induces megabase-scale chromosomal truncations. *Nature communications*, 10(1):1–14, 2019.
- [25] Alexis C Komor, Yongjoo B Kim, Michael S Packer, John A Zuris, and David R Liu. Programmable editing of a target base in genomic dna without double-stranded dna cleavage. *Nature*, 533(7603):420, 2016.
- [26] Y Bill Kim, Alexis C Komor, Jonathan M Levy, Michael S Packer, Kevin T Zhao, and David R Liu. Increasing the genome-targeting scope and precision of base editing with engineered cas9-cytidine deaminase fusions. *Nature biotechnology*, 35(4):371, 2017.
- [27] Keiji Nishida, Takayuki Arazoe, Nozomu Yachie, Satomi Banno, Mika Kakimoto, Mayura Tabata, Masao Mochizuki, Aya Miyabe, Michihiro Araki, Kiyotaka Y Hara, et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science*, 353(6305):aaf8729, 2016.

- [28] Nicole M Gaudelli, Alexis C Komor, Holly A Rees, Michael S Packer, Ahmed H Badran, David I Bryson, and David R Liu. Programmable base editing of a•t to g•c in genomic dna without dna cleavage. *Nature*, 551(7681):464, 2017.
- [29] Seung Woo Cho, Sojung Kim, Yongsub Kim, Jiyeon Kweon, Heon Seok Kim, Sangsu Bae, and Jin-Soo Kim. Analysis of off-target effects of crispr/cas-derived rna-guided endonucleases and nickases. *Genome research*, 24(1):132–141, 2014.
- [30] Yanfang Fu, Jennifer A Foden, Cyd Khayter, Morgan L Maeder, Deepak Reyon, J Keith Joung, and Jeffrey D Sander. High-frequency off-target mutagenesis induced by crispr-cas nucleases in human cells. *Nature biotechnology*, 31(9):822, 2013.
- [31] Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vi-neeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, et al. Dna targeting specificity of rna-guided cas9 nucleases. *Nature biotechnology*, 31(9):827, 2013.
- [32] Tessa G Montague, José M Cruz, James A Gagnon, George M Church, and Eivind Valen. Chopchop: a crispr/cas9 and talen web tool for genome editing. *Nucleic acids research*, 42(W1):W401–W407, 2014.
- [33] Kornel Labun, Tessa G Montague, James A Gagnon, Summer B Thyme, and Eivind Valen. Chopchop v2: a web tool for the next generation of crispr genome engineering. *Nucleic acids research*, 44(W1):W272–W276, 2016.
- [34] Sangsu Bae, Jeongbin Park, and Jin-Soo Kim. Cas-offinder: a fast and versatile algorithm that searches for potential off-target sites of cas9 rna-guided endonucleases. *Bioinformatics*, 30(10):1473–1475, 2014.
- [35] Jeongbin Park, Sangsu Bae, and Jin-Soo Kim. Cas-designer: a web-based tool for choice of crispr-cas9 target sites. *Bioinformatics*, 31(24):4014–4016, 2015.
- [36] Jean-Paul Concordet and Maximilian Haeussler. Crispor: intuitive guide selection for crispr/cas9 genome editing experiments and screens. *Nucleic acids research*, 46(W1):W242–W245, 2018.
- [37] Florian Heigwer, Grainne Kerr, and Michael Boutros. E-crisp: fast crispr target site identification. *Nature methods*, 11(2):122, 2014.
- [38] Venetia Pliatsika and Isidore Rigoutsos. “off-spotter”: very fast and exhaustive enumeration of genomic lookalikes for designing crispr/cas guide rnas. *Biology direct*, 10(1):4, 2015.

- [39] John G Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature biotechnology*, 34(2):184, 2016.
- [40] Yuki Naito, Kimihiro Hino, Hidemasa Bono, and Kumiko Ui-Tei. CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target sites. *Bioinformatics*, 31(7):1120–1123, 2014.
- [41] F Ann Ran, Patrick D Hsu, Chie-Yu Lin, Jonathan S Gootenberg, Silvana Konermann, Alexandro E Trevino, David A Scott, Azusa Inoue, Shogo Matoba, Yi Zhang, et al. Double nicking by RNA-guided CRISPR-Cas9 for enhanced genome editing specificity. *Cell*, 154(6):1380–1389, 2013.
- [42] Prashant Mali, John Aach, P Benjamin Stranges, Kevin M Esvelt, Mark Moosburner, Sriram Kosuri, Luhan Yang, and George M Church. Cas9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature biotechnology*, 31(9):833, 2013.
- [43] Benjamin P Kleinstiver, Vikram Pattanayak, Michelle S Prew, Shengdar Q Tsai, Nhu Nguyen, Zongli Zheng, and J Keith Joung. High-fidelity CRISPR-Cas9 variants with undetectable genome-wide off-targets. *Nature*, 2015.
- [44] Jeongbin Park, Liam Childs, Daesik Kim, Gue-Ho Hwang, Sunghyun Kim, Sang-Tae Kim, Jin-Soo Kim, and Sangsu Bae. Digenome-seq web tool for profiling CRISPR specificity. *Nature methods*, 14(6):548–549, 2017.
- [45] Winston X Yan, Reza Mirzazadeh, Silvano Garnerone, David Scott, Martin W Schneider, Tomasz Kallas, Joaquin Custodio, Erik Wernersson, Yinqing Li, Linyi Gao, et al. Bliss is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nature communications*, 8:15058, 2017.
- [46] Shengdar Q Tsai, Zongli Zheng, Nhu T Nguyen, Matthew Liebers, Ved V Topkar, Vishal Thapar, Nicolas Wyvekens, Cyd Khayter, A John Iafrate, Long P Le, et al. Guide-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature biotechnology*, 33(2):187, 2015.
- [47] Roberto Chiarle, Yu Zhang, Richard L Frock, Susanna M Lewis, Benoit Molinie, Yu-Jui Ho, Darienne R Myers, Vivian W Choi, Mara Compagno, Daniel J Malkin, et al. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell*, 147(1):107–119, 2011.

- [48] Richard L Frock, Jiazhi Hu, Robin M Meyers, Yu-Jui Ho, Erina Kii, and Frederick W Alt. Genome-wide detection of dna double-stranded breaks induced by engineered nucleases. *Nature biotechnology*, 33(2):179, 2015.
- [49] Shengdar Q Tsai, Nhu T Nguyen, Jose Malagon-Lopez, Ved V Topkar, Martin J Aryee, and J Keith Joung. Circle-seq: a highly sensitive in vitro screen for genome-wide crispr-cas9 nuclease off-targets. *Nature methods*, 14(6):607, 2017.
- [50] Daesik Kim, Sangsu Bae, Jeongbin Park, Eunji Kim, Seokjoong Kim, Hye Ryeong Yu, Jinha Hwang, Jong-Il Kim, and Jin-Soo Kim. Digenome-seq: genome-wide profiling of crispr-cas9 off-target effects in human cells. *Nature methods*, 12(3):237, 2015.
- [51] Jeongbin Park and Sangsu Bae. Cpf1-database: web-based genome-wide guide rna library design for gene knockout screens using crispr-cpf1. *Bioinformatics*, 34(6):1077–1079, 2018.
- [52] Jeongbin Park, Kayeong Lim, Jin-Soo Kim, and Sangsu Bae. Cas-analyzer: an online tool for assessing genome editing results using ngs data. *Bioinformatics*, 33(2):286–288, 2017.
- [53] Gue-Ho Hwang, Jeongbin Park, Kayeong Lim, Sunghyun Kim, Jihyeon Yu, Eunchong Yu, Sang-Tae Kim, Roland Eils, Jin-Soo Kim, and Sangsu Bae. Web-based design and analysis tools for crispr base editing. *BMC bioinformatics*, 19(1):542, 2018.
- [54] Kellie A Schaefer, Wen-Hsuan Wu, Diana F Colgan, Stephen H Tsang, Alexander G Bassuk, and Vinit B Mahajan. Unexpected mutations after crispr-cas9 editing in vivo. *Nature methods*, 14(6):547–548, 2017.
- [55] Jeongbin Park, Jin-Soo Kim, and Sangsu Bae. Cas-database: web-based genome-wide guide rna library design for gene knockout screens using crispr-cas9. *Bioinformatics*, 32(13):2017–2023, 2016.
- [56] Tim Hubbard, Daniel Barker, Ewan Birney, Graham Cameron, Yuan Chen, L Clark, Tony Cox, J Cuff, Val Curwen, Thomas Down, et al. The ensembl genome database project. *Nucleic acids research*, 30(1):38–41, 2002.
- [57] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.
- [58] Daesik Kim, Sojung Kim, Sunghyun Kim, Jeongbin Park, and Jin-Soo Kim. Genome-wide target specificities of crispr-cas9 nucleases revealed by multiplex digenome-seq. *Genome research*, 26(3):406–415, 2016.

- [59] Sang-Tae Kim, Jeongbin Park, Daesik Kim, Kyoungmi Kim, Sangsu Bae, Matthias Schlesner, and Jin-Soo Kim. Response to "unexpected mutations after crispr-cas9 editing in vivo". *NATURE METHODS*, 15(4):239–240, 2018.
- [60] Christopher T Saunders, Wendy SW Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012.
- [61] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213, 2013.
- [62] Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen RF Twigg, Andrew OM Wilkie, Gil McVean, Gerton Lunter, WGS500 Consortium, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics*, 46(8):912, 2014.
- [63] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333, 2016.
- [64] Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, 2019.
- [65] Yixue Li and Luonan Chen. Big biological data: challenges and opportunities. *Genomics, proteomics & bioinformatics*, 12(5):187, 2014.
- [66] Susmita Datta, Somnath Datta, Seongho Kim, Sutirtha Chakraborty, and Ryan S Gill. Statistical analyses of next generation sequence data: a partial overview. *Journal of proteomics & bioinformatics*, 3(6):183, 2010.
- [67] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–483, 2010.
- [68] Chang Xu. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and structural biotechnology journal*, 16:15–24, 2018.
- [69] Kary Ocaña and Daniel de Oliveira. Parallel computing in genomic research: advances and applications. *Advances and applications in bioinformatics and chemistry: AABC*, 8:23, 2015.

- [70] Ben Langmead and Abhinav Nellore. Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics*, 19(4):208, 2018.
- [71] Lin Dai, Xin Gao, Yan Guo, Jingfa Xiao, and Zhang Zhang. Bioinformatics clouds for big data manipulation. *Biology direct*, 7(1):43, 2012.
- [72] Jeremy Leipzig. A review of bioinformatic pipeline frameworks. *Briefings in bioinformatics*, 18(3):530–536, 2017.
- [73] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.
- [74] Markus Hsi-Yang Fritz, Rasko Leinonen, Guy Cochrane, and Ewan Birney. Efficient storage of high throughput dna sequencing data using reference-based compression. *Genome research*, 21(5):734–740, 2011.
- [75] Kate Voss, Geraldine Van der Auwera, and Jeff Gentry. Full-stack genomics pipelining with gatk4+ wdl+ cromwell. *F1000Research*, 6, 2017.
- [76] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.
- [77] Enis Afgan, Dannon Baker, Bérénice Batut, Marius Van Den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn A Grüning, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research*, 46(W1):W537–W544, 2018.
- [78] PD Tommaso, EW Floden, C Magis, E Palumbo, and C Notredame. Nextflow, an efficient tool to improve computation numerical stability in genomic analysis. *Biologie aujourd’hui*, 211(3):233–237, 2017.
- [79] Peter Amstutz, Michael R Crusoe, Nebojša Tijanić, Brad Chapman, John Chilton, Michael Heuer, Andrey Kartashov, Dan Leehr, Hervé Ménager, Maya Nedeljkovich, et al. Common workflow language, v1. 0. 2016.
- [80] Philip A Ewels, Alexander Peltzer, Sven Fillinger, Johannes Alneberg, Harshil Patel, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. nf-core: Community curated bioinformatics pipelines. *bioRxiv*, page 610741, 2019.
- [81] Jorrit Boekel, John M Chilton, Ira R Cooke, Peter L Horvatovich, Pratik D Jagtap, Lukas Käll, Janne Lehtiö, Pieter Lukasse, Perry D Moerland, and Timothy J Griffin. Multi-omic data analysis using galaxy. *Nature biotechnology*, 33(2):137–139, 2015.

- [82] Eva Reisinger, Lena Genthner, Jules Kerssemakers, Philip Kensche, Stefan Borufka, Alke Jugold, Andreas Kling, Manuel Prinz, Ingrid Scholz, Gideon Zipprich, et al. Otp: An automatized system for managing and processing ngs data. *Journal of biotechnology*, 261:53–62, 2017.
- [83] Sanjay M Prakadan, Alex K Shalek, and David A Weitz. Scaling by shrinking: empowering single-cell’omics’ with microfluidic devices. *Nature Reviews Genetics*, 18(6):345, 2017.
- [84] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14, 2018.
- [85] Hanna Mendes Levitin, Jinzhou Yuan, and Peter A Sims. Single-cell transcriptomic analysis of tumor heterogeneity. *Trends in cancer*, 4(4):264–268, 2018.
- [86] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 36(5):411–420, June 2018.
- [87] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.
- [88] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017.
- [89] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.
- [90] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O’Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016.
- [91] Joanna S Amberger, Carol A Bocchini, François Schiettecatte, Alan F Scott, and Ada Hamosh. Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1):D789–D798, 2015.

- [92] Julia Richter, Matthias Schlesner, Steve Hoffmann, Markus Kreuz, Ellen Leich, Birgit Burkhardt, Maciej Rosolowski, Ole Ammerpohl, Rabea Wagener, Stephan H Bernhart, et al. Recurrent mutation of the *id3* gene in burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nature genetics*, 44(12):1316, 2012.
- [93] Julia Jabs, Franziska M Zickgraf, Jeongbin Park, Steve Wagner, Xiaoqi Jiang, Katharina Jechow, Kortine Kleinheinz, Umut H Toprak, Marc A Schneider, Michael Meister, et al. Screening drug effects in patient-derived cancer cells links organoid responses to genome alterations. *Molecular systems biology*, 13(11), 2017.
- [94] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic acids research*, 47(D1):D941–D947, 2019.
- [95] Umut Toprak. *Integrative Analysis of Omics Datasets*. PhD thesis, Heidelberg University, 2019.
- [96] Kortine Kleinheinz, Isabell Bludau, Daniel Huebschmann, Michael Heinold, Philip Kenschke, Zuguang Gu, Cristina Lopez, Michael Hummel, Wolfram Klapper, Peter Moeller, et al. Aceseq-allele specific copy number estimation from whole genome sequencing. *BioRxiv*, page 210807, 2017.
- [97] Melinda L Telli, Kirsten M Timms, Julia Reid, Bryan Hennessy, Gordon B Mills, Kristin C Jensen, Zoltan Szallasi, William T Barry, Eric P Winer, Nadine M Tung, et al. Homologous recombination deficiency (hrd) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. *Clinical cancer research*, 22(15):3764–3773, 2016.
- [98] Stephan M Tirier, Jeongbin Park, Friedrich Preußer, Lisa Amrhein, Zuguang Gu, Simon Steiger, Jan-Philipp Mallm, Teresa Krieger, Marcel Waschow, Björn Eismann, et al. pheno-seq—linking visual features and gene expression in 3d cell culture systems. *Scientific reports*, 9(1):1–15, 2019.
- [99] Francesco Pampaloni, Emmanuel G Reynaud, and Ernst HK Stelzer. The third dimension bridges the gap between cell culture and live tissue. *Nature reviews Molecular cell biology*, 8(10):839–845, 2007.
- [100] James T Neal and Calvin J Kuo. Organoids as models for neoplastic transformation. *Annual Review of Pathology: Mechanisms of Disease*, 11:199–220, 2016.

- [101] Hans Clevers. Modeling development and disease with organoids. *Cell*, 165(7):1586–1597, 2016.
- [102] Xin Ye and Robert A Weinberg. Epithelial–mesenchymal plasticity: a central regulator of cancer progression. *Trends in cell biology*, 25(11):675–686, 2015.
- [103] Daniel B Lipka, Tania Witte, Reka Toth, Jing Yang, Manuel Wiesenfarth, Peter Nöllke, Alexandra Fischer, David Brocks, Zuguang Gu, Jeongbin Park, et al. Ras-pathway mutation patterns define epigenetic subclasses in juvenile myelomonocytic leukemia. *Nature communications*, 8(1):1–14, 2017.
- [104] Manuel Rodríguez-Paredes, Felix Bormann, Günter Raddatz, Julian Gutekunst, Carlota Lucena-Porcel, Florian Köhler, Elisabeth Wurzer, Katrin Schmidt, Stefan Gallinat, Horst Wenck, et al. Methylation profiling identifies two subclasses of squamous cell carcinoma related to distinct cells of origin. *Nature communications*, 9(1):1–9, 2018.
- [105] Rabea Wagener, Cristina López, Kortine Kleinheinz, Julia Bausinger, Sietse M Aukema, Inga Nagel, Umut H Toprak, Julian Seufert, Janine Altmüller, Holger Thiele, et al. Igmyc+ neoplasms with precursor b-cell phenotype are molecularly distinct from burkitt lymphomas. *Blood, The Journal of the American Society of Hematology*, 132(21):2280–2285, 2018.
- [106] Teresa G Krieger, Stephan M Tirier, Jeongbin Park, Tanja Eisemann, Heike Peterziel, Peter Angel, Roland Eils, and Christian Conrad. Modeling glioblastoma invasion using human brain organoids and single-cell transcriptomics. *BioRxiv*, page 630202, 2019.
- [107] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.*, 13(4):599–604, April 2018.
- [108] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe’er, Anthony Phillpakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman,

- Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. Science forum: The human cell atlas. *Elife*, 6, December 2017.
- [109] Fredrik Salmén, Sanja Vickovic, Ludvig Larsson, Linnea Stenbeck, Johan Vallon-Christersson, Anna Ehinger, Jari Häkkinen, Åke Borg, Jonas Frisé, Patrik L Ståhl, and Joakim Lundeberg. Multidimensional transcriptomics provides detailed information about immune cell distribution and identity in HER2+ breast tumors. June 2018.
- [110] S Codeluppi, L E Borm, A Zeisel, G La Manno, and others. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods*, 15(11):932–935, 2018.
- [111] Jeffrey R Moffitt, Dhananjay Bambah-Mukku, Stephen W Eichhorn, Eric Vaughn, Karthik Shekhar, Julio D Perez, Nimrod D Rubinstein, Junjie Hao, Aviv Regev, Catherine Dulac, and Xiaowei Zhuang. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416), November 2018.
- [112] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, April 2015.
- [113] Eric Lubeck, Ahmet F Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods*, 11(4):360–361, April 2014.
- [114] Rongqin Ke, Marco Mignardi, Alexandra Pacureanu, Jessica Svedlund, Johan Botling, Carolina Wählby, and Mats Nilsson. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods*, 10(9):857–860, September 2013.
- [115] Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C Ferrante, Richard Terry, Brian M Turczyk, Joyce L Yang, Ho Suk Lee, John Aach, Kun Zhang, and George M Church. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.*, 10(3):442–458, March 2015.
- [116] Silas Maniatis, Tarmo Äijö, Sanja Vickovic, Catherine Braine, Kristy Kang, Annelie Mollbrink, Delphine Fagegaltier, Žaneta Andrusivová, Sami Saarenpää, Gonzalo Saiz-Castro, Miguel Cuevas, Aaron Watters, Joakim Lundeberg, Richard Bonneau, and Hemali Phatnani. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science*, 364(6435):89–93, April 2019.
- [117] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg,

- Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisé. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, July 2016.
- [118] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernández Navarro, Joshua Gould, Gabriel K Griffin, Åke Borg, Mostafa Ronaghi, Jonas Frisé, Joakim Lundeberg, Aviv Regev, and Patrik L Ståhl. High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods*, 16(10):987–990, October 2019.
- [119] Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400), July 2018.
- [120] Jeffrey M Perkel. Starfish enterprise: finding RNA patterns in single cells. *Nature*, 572(7770):549–551, August 2019.
- [121] Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, 92(2):342–357, October 2016.
- [122] Keren Bahar Halpern, Rom Shenhav, Orit Matcovitch-Natan, Beata Toth, Doron Lemze, Matan Golan, Efi E Massasa, Shaked Baydatch, Shanie Landen, Andreas E Moor, Alexander Brandis, Amir Giladi, Avigail Stokar Avihail, Eyal David, Ido Amit, and Shalev Itzkovitz. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*, 542(7641):352–356, February 2017.
- [123] Jocelyn Y Kishi, Brian J Beliveau, Sylvain W Lapan, Emma R West, Allen Zhu, Hiroshi M Sasaki, Sinem K Saka, Yu Wang, Constance L Cepko, and Peng Yin. SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. *Nat. Methods*, 16(6):533–544, May 2019.
- [124] Antti Lignell, Laura Kerosuo, Sebastian J Streichan, Long Cai, and Marianne E Bronner. Identification of a neural crest stem cell niche by spatial genomic analysis. *Nat. Commun.*, 8(1):1830, November 2017.
- [125] Rintu Maria Thomas and Jisha John. A review on cell detection and segmentation in microscopic images. *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*, pages 1–5, 2017.

- [126] Emanuel Parzen. On estimation of a probability density function and mode. *Ann. Math. Stat.*, 33(3):1065–1076, 1962.
- [127] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.*, 27(3):832–837, 1956.
- [128] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. March 2019.
- [129] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, March 2015.
- [130] Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job van der Zwan, Martin Häring, Emelie Braun, Lars E Borm, Gioele La Manno, Simone Codeluppi, Alessandro Furlan, Kawai Lee, Nathan Skene, Kenneth D Harris, Jens Hjerling-Leffler, Ernest Arenas, Patrik Ernfors, Ulrika Marklund, and Sten Linnarsson. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014.e22, August 2018.
- [131] Sueli Marques, Amit Zeisel, Simone Codeluppi, David van Bruggen, Ana Mendanha Falcão, Lin Xiao, Huiliang Li, Martin Häring, Hannah Hochgerner, Roman A Romanov, Daniel Gyllborg, Ana Muñoz Manchado, Gioele La Manno, Peter Lönnerberg, Elisa M Floriddia, Fatemah Rezayee, Patrik Ernfors, Ernest Arenas, Jens Hjerling-Leffler, Tibor Harkany, William D Richardson, Sten Linnarsson, and Gonçalo Castelo-Branco. Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science*, 352(6291):1326–1329, June 2016.
- [132] Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, Osnat Penn, Trygve Bakken, Vilas Menon, Jeremy Miller, Olivia Fong, Karla E Hirokawa, Kanan Lathia, Christine Rimorin, Michael Tieu, Rachael Larsen, Tamara Casper, Eliza Barkan, Matthew Kroll, Sheana Parry, Nadiya V Shapovalova, Daniel Hirschstein, Julie Pendergraft, Heather A Sullivan, Tae Kyung Kim, Aaron Szafer, Nick Dee, Peter Groblewski, Ian Wickersham, Ali Cetin, Julie A Harris, Boaz P Levi, Susan M Sunkin, Linda Madisen, Tanya L Daigle, Loren Looger, Amy Bernard, John Phillips, Ed Lein, Michael Hawrylycz, Karel Svoboda, Allan R Jones, Christof

- Koch, and Hongkui Zeng. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, November 2018.
- [133] Dmitry Gerashchenko, Jonathan P Wisor, Deirdre Burns, Rebecca K Reh, Priyattam J Shiromani, Takeshi Sakurai, Horacio O de la Iglesia, and Thomas S Kilduff. Identification of a population of sleep-active cerebral cortex neurons. *Proc. Natl. Acad. Sci. U. S. A.*, 105(29):10227–10232, July 2008.
- [134] Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, Darren Bertagnolli, Jeff Goldy, Nadiya Shapovalova, Sheana Parry, Changkyu Lee, Kimberly Smith, Amy Bernard, Linda Madisen, Susan M Sunkin, Michael Hawrylycz, Christof Koch, and Hongkui Zeng. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.*, 19(2):335–346, February 2016.
- [135] Ryohei Tomioka, Keiko Okamoto, Takahiro Furuta, Fumino Fujiyama, Takuji Iwasato, Yuchio Yanagawa, Kunihiro Obata, Takeshi Kaneko, and Nobuaki Tamamaki. Demonstration of long-range GABAergic connections distributed throughout the mouse neocortex. *Eur. J. Neurosci.*, 21(6):1587–1600, March 2005.
- [136] Vincent A Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.
- [137] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nat. Methods*, 10(11):1127–1133, November 2013.
- [138] Ryan A Flynn, Benjamin A H Smith, Alex G Johnson, Kayvon Pedram, Benson M George, Stacy A Malaker, Karim Majzoub, Jan E Carette, and Carolyn R Bertozzi. Mammalian Y RNAs are modified at discrete guanosine residues with n-glycans. September 2019.
- [139] Xiaoyan Qian, Kenneth D. Harris, Thomas Hauling, Dimitris Nicoloutsopoulos, Ana B. Muñoz-Manchado, Nathan Skene, Jens Hjerling-Leffler, and Mats Nilsson. A spatial atlas of inhibitory cell types in mouse hippocampus. *bioRxiv*, 2018.
- [140] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, 2008.
- [141] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32(3):29, June 2013.

- [142] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, pages 61–70, 2006.
- [143] Michael Garland and Paul S Heckbert. Surface simplification using quadric error metrics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 209–216, 1997.
- [144] H Hoppe. New quadric metric for simplifying meshes with appearance attributes. In *Proceedings Visualization '99 (Cat. No.99CB37067)*, pages 59–510, October 1999.
- [145] Will Schroeder, Ken Martin, and Bill Lorensen. *The Visualization Toolkit: An Object-oriented Approach to 3D Graphics*. Kitware, 2006.
- [146] Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, and Guido Ranzuglia. Meshlab: an open-source mesh processing tool. In *Eurographics Italian chapter conference*, volume 2008, pages 129–136, 2008.
- [147] Rebecca D Hodge, Trygve E Bakken, Jeremy A Miller, Kimberly A Smith, Eliza R Barkan, Lucas T Graybuck, Jennie L Close, Brian Long, Nelson Johansen, Osnat Penn, Zizhen Yao, Jeroen Eggermont, Thomas Höllt, Boaz P Levi, Soraya I Shehata, Brian Aevermann, Allison Beller, Darren Bertagnolli, Krissy Brouner, Tamara Casper, Charles Cobbs, Rachel Dalley, Nick Dee, Song-Lin Ding, Richard G Ellenbogen, Olivia Fong, Emma Garren, Jeff Goldy, Ryder P Gwinn, Daniel Hirschstein, C Dirk Keene, Mohamed Keshk, Andrew L Ko, Kanan Lathia, Ahmed Mahfouz, Zoe Maltzer, Medea McGraw, Thuc Nghi Nguyen, Julie Nyhus, Jeffrey G Ojemann, Aaron Oldre, Sheana Parry, Shannon Reynolds, Christine Rimorin, Nadiya V Shapovalova, Saroja Somasundaram, Aaron Szafer, Elliot R Thomsen, Michael Tieu, Gerald Quon, Richard H Scheuermann, Rafael Yuste, Susan M Sunkin, Boudewijn Lelieveldt, David Feng, Lydia Ng, Amy Bernard, Michael Hawrylycz, John W Phillips, Bosiljka Tasic, Hongkui Zeng, Allan R Jones, Christof Koch, and Ed S Lein. Conserved cell types with divergent features in human versus mouse cortex. *Nature*, 573(7772):61–68, September 2019.
- [148] Thomas A Caswell, Michael Droettboom, John Hunter, Antony Lee, Eric Firing, David Stansby, Jody Klymak, Elliott Sales de Andrade, Jens Hedegaard Nielsen, Nelle Varoquaux, Tim Hoffmann, Benjamin Root, Phil Elson, Ryan May, Darren Dale, Jae-Joon Lee, Jouni K Seppänen, Damon McDougall, Andrew Straw, Paul Hobson, Christoph Gohlke, Tony S Yu, Eric Ma, Adrien F Vincent, Steven Silvester, Charlie Moad, Jan Katins, Nikita Kniazev, Federico Ariza, and Elan Ernest. `matplotlib/matplotlib: REL: v3.1.1`, July 2019.

- [149] Michael Waskom, Olga Botvinnik, Drew O’Kane, Paul Hobson, Joel Ostblom, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B Cole, Jordi Warmenhoven, Julian de Ruiters, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Thomas Brunner, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, and Adel Qalieh. `mwaskom/seaborn: v0.9.0` (july 2018), July 2018.

Acknowledgements

제일 먼저, 다사다난했던 박사 과정 기간 동안 저를 지도해 주신 Matthias Schlesner 박사님 및 Naveed Ishaque 박사님, 그리고 Roland Eils 교수님, Benedikt Brors 교수님께 감사를 드립니다.

(First of all, I would like to give thanks to my daily supervisors Matthias Schlesner and Naveed Ishaque, and my university supervisors Roland Eils and Benedikt Brors.)

그리고 함께 SSAM 을 개발한 최원일 군, 그리고 RGEN 관련 툴을 함께 만든 황규호 학생, 유은총 형, 또 마지막 논문 마무리에 도움을 준 Sebastian Tiesmeyer, Luca Tosti, Bianca Hennig, 그 외 다른 프로젝트로 함께 일했던 Julia Jabs, Stephan Tirier, 또 컴퓨터 셋업 및 클러스터 관리를 도와 준 Rolf Kabbe, Alexander Balz, Frank Thommen, Martin Lang, Karlheinz Groß, Hans Bartelmeß, 또 pipeline 제작 및 분석에 도움을 준 Michael Heinold, Philip Reiner Kensche, 또한 행정 처리 및 문서 작성 등의 일들을 도와 주셨던 Corinna Sprengart, Manuela Schäfer, Alexandra Friedrich, Franziska Gudrun Müller, Sandra Bodogh, Jan Eufinger, 그리고 연구실에서 함께 생활하며 동고동락했던 동료들, Daniel Hübschmann, Stephen Krämer, Gregor Warsow, Zuguang Gu, Jing Yang, Ling Hai, Dina Krämer, Dina ElHarouni, Julian Seufert, 특히 졸업논문 교정에 많은 도움을 준 Nagarajan Paramasivam, Dorett Odoni, 그리고 배상수 선배에게 큰 감사와 고마움을 전합니다.

(And big thanks to Wonyl Choi who is the other main developer of SSAM, Gue-Ho Hwang and Eunchong Yu for the co-development of the some RGEN tools, Sebastian Tiesmeyer, Luca Tosti, and Bianca Hennig for their help with finalizing the paper manuscript. And also thanks to Julia Jabs and Stephan Tirier, who inspired me to develop dedicated bioinformatics pipelines. I also would like to mention the help from our system administrative staffs, Rolf Kabbe, Alexander Balz, Frank Thommen, Martin Lang for the maintenance of cluster system, Karlheinz Groß, Hans Bartelmeß for the hardware management, and Michael Heinold and Philip Reiner Kensche for the help with the pipeline development. Also thanks to Corinna Sprengart, Manuela Schäfer, Alexandra Friedrich, Franziska Gudrun Müller, Sandra Bodogh, Jan Eufinger for the help regarding administrative stuffs, and to my colleagues and friends, Daniel Hübschmann, Stephen Krämer, Gregor Warsow, Zuguang Gu, Jing Yang, Ling Hai, Dina Krämer, Dina ElHarouni, and Julian Seufert (especially for my residence permit problems!). Especially I really appreciate the help from Nagarajan Paramasivam, Dorett Odoni, and

Sangsu Bae, for the review and correction of my thesis.)

특히 데이터를 아무 조건없이 제공해 주고 또 최종 결과 분석에 도움을 준 분들, Lars Borm, Simone Codeluppi, Jeffery Moffit, Yue (Tony) Zhuo, Brian Long, Emma Garren, Thuc Nghi Nguyen, Bosiljka Tasic, Jeremy Miller, Ambrose Carr, Ed Lein에게 다시 한번 감사를 전합니다. 이들의 도움이 없었다면 저의 졸업은 불가능했을 것입니다.

(Special thanks to the data providers and the ones who helped data analysis, Lars Borm, Simone Codeluppi, Jeffery Moffit, Yue (Tony) Zhuo, Brian Long, Emma Garren, Thuc Nghi Nguyen, Bosiljka Tasic, Jeremy Miller, Ambrose Carr, and Ed Lein. Without their help, my graduation would not have been possible.)

그리고 아주대 초등과학반 친구들 권진규, 김기범, 김석영, 배주영, 이지현, 조병주, 추준호, 고승희, (요즘은 연락이 잘 안 되는) 유지수, 그리고 물리학과 08학번 친구들 이유정, 이주환, 정하나, 차민령에게, 그리고 석사 동기 박민호 군에게, 또 독일에서 공부하느라 같이 고생하는 김광영 군, 또 근래 도움을 많이 주신 이상화 선배, 그리고 KIST 정철현 박사님께도 고마움을 전합니다.

마지막으로 항상 저를 지켜봐 주시고 도와주시는 사랑하는 부모님 및 가족들, 동생 박혜연, 그리고 지영과 지영 가족들에게, 모두모두 감사를 전합니다. 사랑합니다.