# Deconstructing spinal interneurons, one cell type at a time

Mariano Ignacio Gabitto

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

# Deconstructing spinal interneurons, one cell type at a time

Mariano Ignacio Gabitto

## Abstract

Documenting the extent of cellular diversity is a critical step in defining the functional organization of the nervous system. In this context, we sought to develop statistical methods capable of revealing underlying cellular diversity given incomplete data sampling - a common problem in biological systems, where complete descriptions of cellular characteristics are rarely available. We devised a sparse Bayesian framework that infers cell type diversity from partial or incomplete transcription factor expression data. This framework appropriately handles estimation uncertainty, can incorporate multiple cellular characteristics, and can be used to optimize experimental design. We applied this framework to characterize a cardinal inhibitory population in the spinal cord.

Animals generate movement by engaging spinal circuits that direct precise sequences of muscle contraction, but the identity and organizational logic of local interneurons that lie at the core of these circuits remain unresolved. By using our Sparse Bayesian approach, we showed that V1 interneurons, a major inhibitory population that controls motor output, fractionate into diverse subsets on the basis of the expression of nineteen transcription factors. Transcriptionally defined subsets exhibit highly structured spatial distributions with mediolateral and dorsoventral positional biases. These distinctions in settling position are largely predictive of patterns of input from sensory and motor neurons, arguing that settling position is a determinant of inhibitory

microcircuit organization.  Finally, we extensively validated inferred cell types by direct experimental measurement and then, extend our Bayesian framework to full transcriptome technologies. Together, these findings provide insight into the diversity and organizational logic through which inhibitory microcircuits shape motor output.

# Contents

# List of Figures

# List of Tables

# Acknowledgements

This thesis would not have been possible without the support of a huge number of people. I would like to greatly thank my graduate adviser, Charles Zuker, for his support; the amazing journey in which I embarked would have not been possible without the passion and audacity with which Charles approaches science. Every step of the journey was a magnificent learning experience.

I am grateful to Thomas Jessell, Liam Paninski and Larry Abbott for their guidance and their teachings. I learned a lot from them and this, helped me grow as a scientist. I will treasure the discussions in each one offices tackling science from different points of view. This work would not have been possible without instrumental contributions from Jay Bikoff and Ari Pakman, who were big allies in tackling the inordinate diversity of spinal interneurons. I was really lucky to find collaborators of such a caliber.

My thesis committee, John Cunningham, Richard Axel and Tom Maniatis generously gave me their time and advice. Richard was essential in providing scientific guidance and common sense about this thesis and Tom was essential in keeping common sense and Charles at bay, on top of providing encouragement and scientific advice. I am also grateful to my external committee member, Yaniv Erlich, for being willing to read my thesis and take part in the defense. In addition, Nate Sawtell, provided scientific guidance, his time and advice. Electrophysiological experiments were performed in the laboratory of Paco Alvarez at Emory, I am grateful to Paco for sharing the data and advice on analysis.

For consistently excellent advice, as well as for generously sharing their knowledge, I would like to thank many members of the Zuker lab. In particular, I would like to thank Xiaoke Chen, Jayaram Chandrashekar, Lindsey McPherson and Hojoon Lee for their guidance and scientific training. I would like to thank Martina Car and Ariel Suazo-Maler for invaluable technical assistant and Laura Rickman for keeping the lab running. Martin Vignovich, Alex Sisti, Hershy Fishman and Ryan Lessard gave advice and were always a source of good feedback and incredible lab mates. In addition, I would like to thank members of the Axel lab, Felicity Gore, Edmund Schwarts and members of the Jessell lab Thomas Reardon, Andy Murray, David Ng for invaluable technical assistance, instruction and, advice in all aspects of viral vectors generation. I would like to thank many members of the Maniatis lab who were important for keeping me company during many nights at the lab, Juan Carlos Tapia, Yiorgos Mountoufaris, Daniele Cancio and Noam Rudnick. Monica Carrasco taught me all I know about cell culture and Juan Carlos Tapia provided consistent advice on microscopy. Both were an incredible source of positive feedback.

The program directors of the Neurobiology and Behavior program: Darcy Kelley, Carol Mason, and Ken Miller were a source of guidance. And our departmental administrators Alla Kerzhner, Cecil Oberbeck and Rozzana Yacub ensured that I continued on track every semester in the program as a consistently enrolled, student and were instrumental in setting up my defense.

I would also like to thank my friends, David Pfau, Armen Enikolopov, Daniel Morozov, Shobhit Singla, Rebecca Brachman, Diego Zimet, Diego Gutnisky, Brian Wundheiler and Tim Machado. They were a constant source of feedback and support. I would like to thank my mother for everything over the years. Last but not least, Cris, accompany me along the years with love and care, I learned from her to never surrender.

Materials presented in this dissertation have been included in the publications below. Permission was obtained from the publishers for inclusion.

Gabitto, M.I., Pakman, A., Bikoff, J.B., Abbott, L.F., Jessell, T.M., and Paninski, L. (2016). Bayesian sparse regression analysis documents the diversity of spinal inhibitory interneurons. Cell.

Bikoff J.B., Gabitto M.I., Rivard A.F., Drobac E., Machado T.A., Miri A., Brenner-Morton S., Famojure E., Diaz C., Alvarez F.J., et al. (2016). Spinal inhibitory interneuron diversity delineates variant motor microcircuits. Cell.

# Chapter 1

# Introduction

## 1.1 General Overview

The primary function of the nervous system is to integrate information about the external world to produce behavior. This response translates into the coordinate activation of muscles by circuits residing in the spinal cord. But, before we can attain the study of the neuronal circuits at play when performing movement, we must be able to deconstruct such circuits into their neuronal types. At present, there is no scientific consensus on what a neuronal cell type is, since neurons can be distinguished by many factors e.g. their patterns of gene expression, neurotransmitters, electrophysiological properties, morphology, connectivity. Nonetheless, a provisional census can be assembled by characterizing intrinsic neuronal properties, which in turn enables the manipulation of cell types in controlled ways and, the understanding of how these types change during disease.

In this thesis, I will focus on two main objectives: elucidating the organization of a major interneuronal population in the spinal cord implicated in the regulation of locomotor activity and, at the same time, setting forth a statistical framework for characterizing cellular populations. Such a framework will inform us about the analysis of sparsely sampled datasets, a commonly encountered problem in the biological sciences. To resolve this issue, we resorted to the assignment of uncertainty to the confidence on the existence of each individual cell type.

The spinal cord motor system, the unit of the CNS in charge of producing movement, presents many advantages in terms of characterizing its cellular constituents: its output is well characterize and divided into labelled lines consisting of sets of motor neurons innervating different muscles; the developmental program of motor neurons has been extensively studied and delineated [Jessell, 2000]; the physiology and connectivity of its neurons have been clearly outlined at the mature stage [Brown 1981]; its cells are assembled in a stereotypical manner; the circuits implied in motor control are largely confined to the ventral region; and much work has been done in molecularly characterizing their local and output neurons. For these reasons, the spinal cord can be used as a model for understanding how neuronal populations are distinguished while teaching us about the most essential neuronal circuits involved in movement.

Within the spinal cord, local interneurons have major roles in shaping the activity of motor output pathways and amongst these cells, the ones that release inhibitory transmitters critically affect motor neuron output. Spinal interneurons involved in the generation of locomotor activity reside in the ventral region of the cord and, they originate during development from four progenitor domains, termed V0 – V3. To gain insight into the functional organization of local inhibitory circuits for motor control we have focused on diversity and connectivity within the cardinal interneuronal population, V1. To characterize V1, we first identified transcription factors that can help to fractionate this parental population. We focused on transcription factors given their preponderant role during development in assigning cellular identity in the nervous system [Jessell, 2000]; [Arber 2012]; [Goulding et al, 2002]. We then measured first and second order statistics – the fraction of cells expressing and co-expressing each protein within V1– by immunohistochemistry. This methodology provides additional information regarding cell location, a cell type characteristic important for determining connectivity and functional cell properties.

Using these datasets, we employed our statistical framework to infer underlying cell types within V1, highlighting the differences and advantages of using more sources of cellular characterization to constrain inference procedures. Next, we validated our inferred cell types by further immunohistochemical analysis, spatial location, and quantitative PCR experiments. We then examined the functional consequences of such diversity by electrophysiological recordings. Finally, we studied how interneuronal settling position relates to patterns of input from sensory and motor neurons, with a focus on differential wiring from inhibitory microcircuits into motor pools controlling hip, ankle, and foot muscles.

## 1.2 Cellular diversity

Throughout the animal kingdom, tissues and organs are comprised of diverse cell types, each possessing a characteristic morphology and specialized function. In mammals, the diversification of cell types attains its most impressive extremes in the hematopoietic and nervous systems. The hematopoietic system contains more than two hundred distinct cell types, the individual identities of which are initially determined by the selective expression of DNA-binding transcription factors [Jojic et al, 2013]. In the mammalian central nervous system (CNS), however, the extent of neuronal diversity has yet to be resolved [Insel et al, 2013]. In part this reflects the fact that criteria for distinguishing one neuronal type from another remain tenuous. Classification schemes based on connectivity, molecular profile, or physiology at best capture only isolated features of neuronal heterogeneity [Masland, 2004]; [Defelipe et al, 2013]; [Sharpee, 2014]. A classification method is needed that identifies cell types using all these different cellular features.

At a molecular level, distinctions in neuronal cell type originate during embryogenesis, and depend on the activities of transcription factors that promote the expression of downstream effectors [Kohwi and Doe, 2013]. Attempts to define the link between transcriptional identity and neuronal diversity have focused mainly on the analysis of long-distance projection neurons, for which distinctions in target innervation provide a clear correlate of functional divergence [Molyneaux et al, 2007]; [Sanes and Masland, 2015]. Thus in the retina and cerebral cortex, different functional classes of ganglion and pyramidal neurons have been delineated through their transcriptional identities [Siegert et al, 2009]; [Greig et al, 2013]; [Ascoli et al, 2008]; [Fishell and Rudy, 2011). Similarly, in the spinal cord the hierarchical ordering of motor neuron subtypes and their connections is known to have its origins in discrete profiles of transcription factor expression [Ericsson et al, 1996]; Dasen et al, 2005]; [Dasen and Jessell, 2009].

Yet local interneurons represent the most prevalent neuronal class within the mammalian CNS, collectively shaping the output of long-range-projection neurons [Isaacson and Scanziani, 2011]. Among local neurons, cortical inhibitory interneurons are the most widely studied interneuronal types in the CNS [DeFelipe et al., 1989]; [Hendry et al., 1989]; [Fonseca et al., 1993]; [Gabbott and Bacon, 1996]. They can be molecularly identified based on the expression of several markers: *parvalbumin*, found in chandelier and fast spiking basket, or wide arbor, cells; *calretinin* [Cond´e et al., 1994]; [Gabbott and Bacon, 1996], labelling double bouquet cells, Cajal-Retzius cells, and bipolar cells; *calbindin* [Hendry et al., 1989]; [DeFelipe et al., 1990]; [Kubota et al., 1994]; [del R´ıo and DeFelipe, 1995]; [Gabbott and Bacon, 1996], marking double bouquet cells; and *somatostatin* [Kubota et al., 1994]; [Kawaguchi and Kubota, 1996]; [Smiley et al., 2000], which labels Martinotti cells (Figure 1.2a). Nonetheless, recent work have demonstrated that these classes of inhibitory interneurons can be further subdivide, establishing the premise that various interneuron classes are not ultimately defined by the expression of a single gene; Other morphological features must be probed in the end when determining the identity of cell types ([Zeisel 2015] and [Cortical Allen Brain 2016]; [Kepecs & Fishell]; Figure 1.2b-c).

Resolving the extent of interneuron diversity within the spinal cord has remained a challenge, not least because the detailed organization of spinal circuits and their cellular populations is yet to be uncovered.

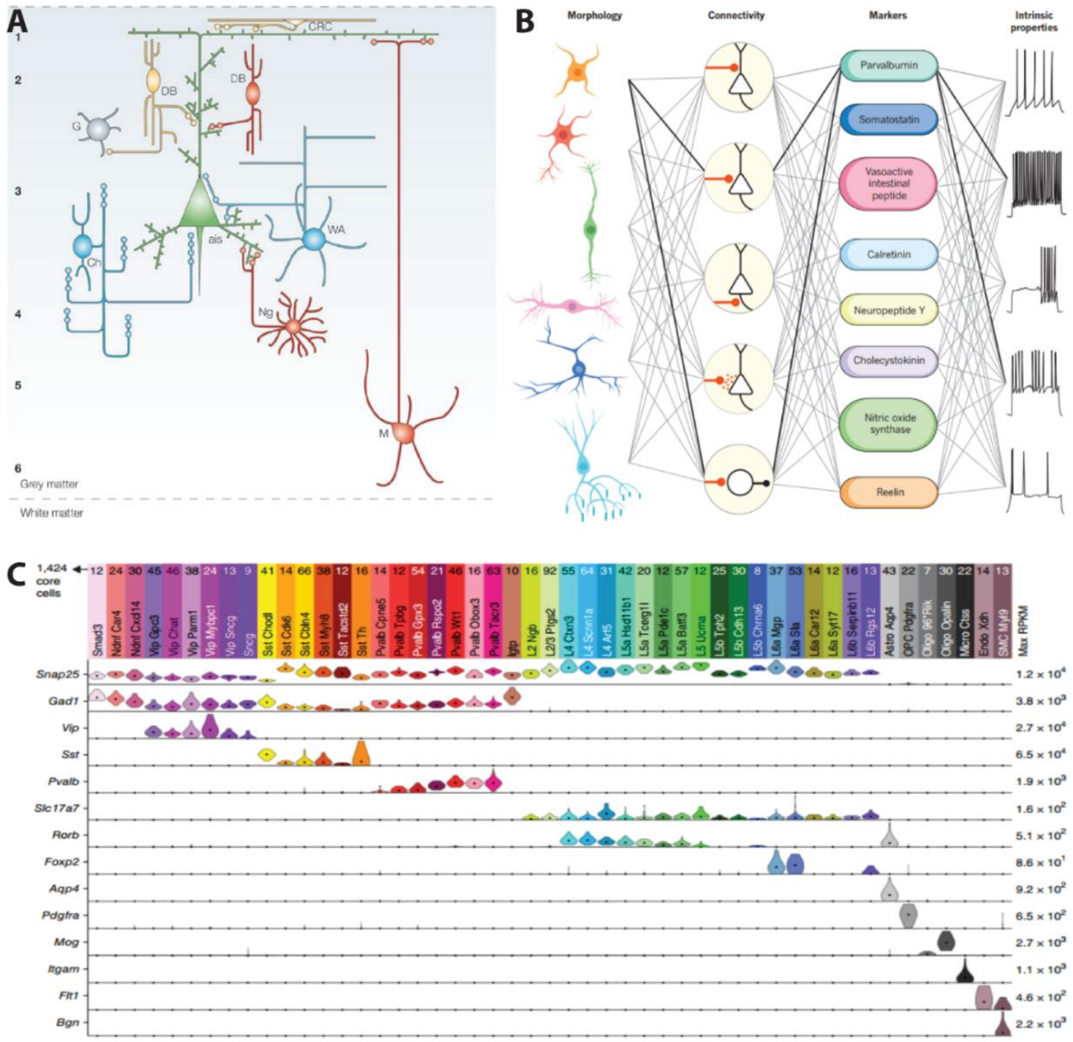


**Figure 1.1. Cortical inhibitory interneurons are heterogeneous, with various subtypes distinguished by different morphological, physiological and molecular characteristics.**

(a) Diagram depicting distinct subpopulations of GABA neurons in prefrontal cortex of humans known to have specialized functions in regulating pyramidal neuron activity. Color code indicates widely used markers distinguishing different types: parvalbumin (blue), calbindin (red), and calretinin (yellow). Chandelier (Ch) cells, and wide arbor (WA) or basket cells both parvalbumin expressing neurons. Neurogliaform (Ng), Martinotti (M), and double bouquet (DB, red) cells express calbindin. Double bouquet (DB, yellow) cells express calretinin. Reproduced from Lewis et al. (2005).

(b) Multiple dimensions of interneuron diversity can be used to define each cell type. Typically a combination of morphological, connectivity pattern, synaptic properties, marker expression and intrinsic firing define the properties of each type (Diagram reproduced from Kepecs and Fishell, 2014).

(c) 49 transcriptomic cell types, including 23 GABAergic, 19 glutamatergic and 7 non-neuronal types identified by Tasic et al, 2016. Violin plots characterize distribution of mRNA expression on a linear scale for already known marker genes (Reproduced from [Tasic et al, 2016]).

## 1.3 The organization of spinal circuits

The rich repertoire of movements produced by the nervous system is generated by numerous neuronal circuits that at last lead to the precise and coordinated activation of muscles. At the core of the motor control system lies the spinal cord, collecting and integrating diverse sources of information; its ultimate task is to drive muscle contraction through its output cells, motor neurons. One of the earliest attempts to order the cellular architecture of the spinal cord that remains widely used, was devised by Rexed, who subdivided the cord into ten rostro-caudal laminae that imposes a consistent and inclusive cytoarchitectonic scheme [Rexed, 1952]. The delimitation of these layers and regions is thus based primarily on the cytological characteristics of the nerve cells and their cytoarchitectonic aggregations (Figure 1.2a).

At that time, Romanes was pursuing a different characterization striving to identify the functional significance of the various motor cell groups within the ventral horn of the spinal cord, which at the time were unknown (for a review of Romanes work, see [Jessell et al, 2011]). In two seminal papers, he established that the ventral region of the spinal cord, Lamina IX, contains sets of motor neurons arrayed in longitudinal columns projecting to distinct muscles in the periphery [Romanes, 1951]; [Romanes, 1964]; (Figure 1.2b). These findings were later confirmed with exquisite detail for the lumbosacral ventral motor neurons [Vanderhorst and Holstege, 1997]. The arrangement of motor neurons into lateral motor columns revealed a multilayered topographic organization that links pool identity and location of motor neurons to the spatial arrangement and biomechanical features of limb muscles. Next, motor neurons receive local inputs from interneurons positioned in the ventral spinal cord. These cells represents the following step in the motor processing hierarchy.

Independent of the behavior enacted, spinal interneurons participate in a three component system mediating movement (Figure 1.2c). First, supraspinal centers in brainstem and higher brain areas send descending input into the cord [Orlovsky et al., 1999]; [Grillner et al., 2005]. The planning of actions prior to overt signs of movement are thought to be in part the work of circuits in cortical centers and other supraspinal circuits; For instance, primary motor cortex sends descending commands into the spinal cord, where they are involved in the control of many voluntary motor programs [Miri et al, 2013]. Second, cutaneous, proprioceptive, and nociceptive sensory systems, that constantly monitor the consequences of motor actions [Brown, 1981]; [Rossignol et al., 2006]; [Windhorst, 2007] are assigned the job of conveying information about the state of muscle activation to central neurons, sending afferents onto spinal interneurons located in the first five laminae within the dorsal horn [Lallemend & Ernfors, 2012]; Figure 1.2d). Third, local interneurons, in between sensory and motor neurons, play an important role in shaping motor output. Motor neurons residing on Lamina IX of the ventral horn mainly receive inputs from interneurons positioned in the ventral region of the spinal cord [Tripodi et al., 2011; Kjaerulff & Kiehn, 1996]. Ventral interneurons are essential to generate rhythmic activity [Orlovsky et al., 1999]; [Jankowska, 2001]; [Kiehn, 2011], they influence left-right alternation, flexor-extensor patterning, speed of rhythmic motor output, and burst robustness [Arber, 2012]. The interaction between these three components of the motor hierarchy provides the plasticity and behavioral diversification observed in nature. This diversification occurs through orchestrated interactions of transcription factors regulating differentiation during development.

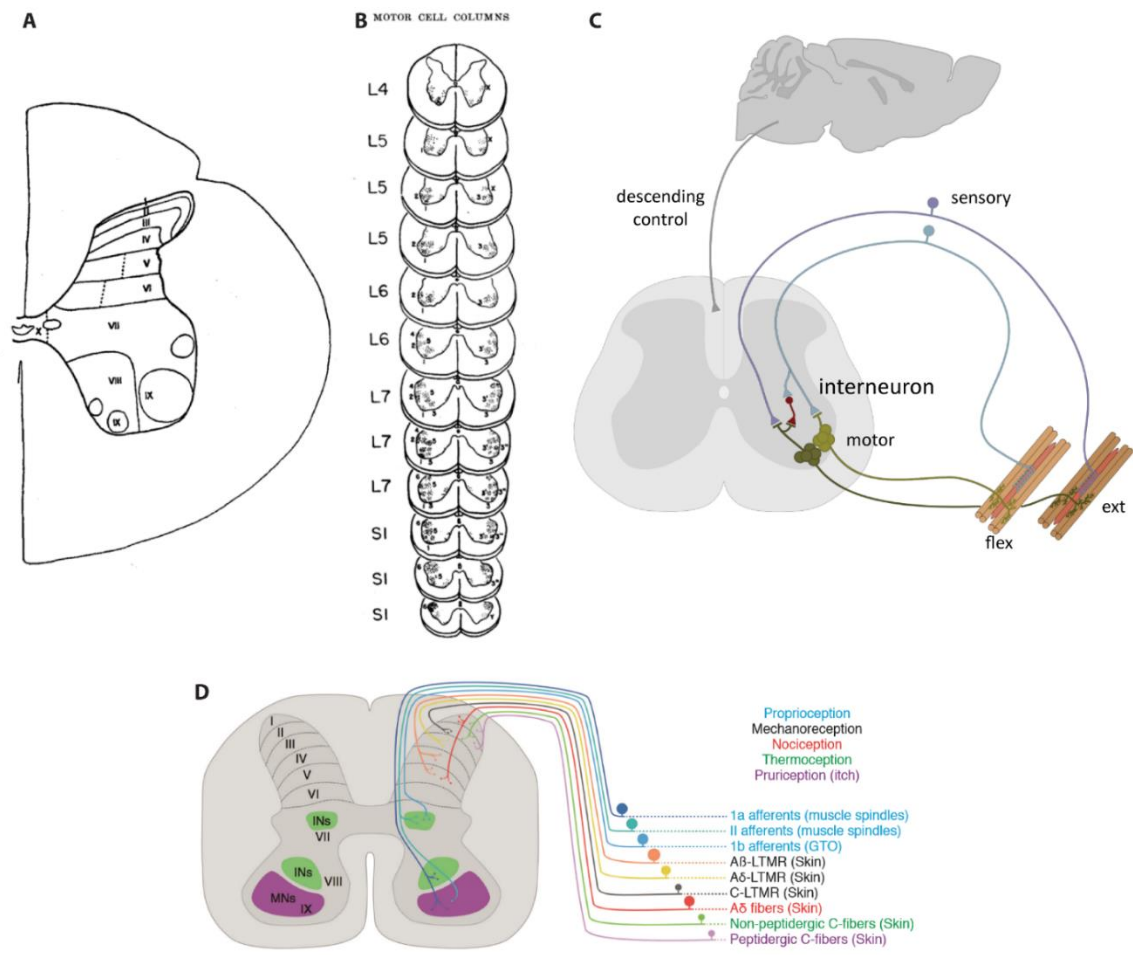


**Figure 1.2. The organization of spinal circuits.**

(a) Schematic drawing of the laminar organization of the spinal cord gray matter of the fifth lumbar segment in the adult cat as proposed by Rexed (adapted from [Rexed, 1952]).
(b) The motor cell columns in cat. Each number depicts individual columns along the different lumbosacral sections (adapted from [Romanes, 1951]).
(c) Schematic diagram depicting interneuron prominent position within spinal circuits, integrating descending and sensory inputs to coordinate motor neurons activation.
(d) Axonal termination patterns onto the different spinal laminae from sensory neurons in dorsal root ganglia. Group Ia proprioceptive afferents and some group II afferents, which innervate muscle spindles, respond to muscle stretch in the periphery. Group Ib fibers, which innervate the Golgi tendon organs, respond to muscle tension. Group II proprioceptors, which innervate secondary endings in muscle spindles (adapted from [Lallemend and Ernfors, 2012]).

## 1.4 Molecular diversification of spinal neurons

Spinal circuits tasked to control limb movement depend on interneurons to drive motor neurons in precisely organized patterns. Furthermore, interneurons serve as relays for sensory feedback and descending command pathways [Schomburg, 1990]; [Windhorst, 2007]; [Arber, 2012]. Such varied requirements, in addition to the large number of muscles to be controlled, demand an inordinate degree of motor and interneuronal heterogeneity necessary to accommodate the diverse functions of spinal microcircuits.

Diversification of spinal neurons has its origins at early stages of development mediated by a secreted protein, Sonic Hedgehog (Shh), which differentiates cells into an array of spatially restricted progenitors along the dorsoventral axis during temporally restricted periods [Jessell, 2000] (Figure 1.3a). Even at this early stage, position and molecular distinctions occupy a privileged role in defining neuronal identity in the spinal cord. Shh signaling, in concert with a transcription factor code, divides neurons into six dorsal and five ventral cardinal populations [Alaynick et al., 2011]; [Goulding, 2009]; [Jessell, 2000]; [Kiehn, 2011]; [Dasen & Jessell, 2009]. These cardinal interneuronal populations are termed V0-V3 and dI1-dI6 (ventral and dorsal domains respectively) as well as motor neurons (Figure 1.3b). Although much progress has been done in defining the cellular types arising from these cardinal domains, we still do not known the complete transcription factor profile of any single class of spinal interneurons.

How many interneuronal populations are needed to produce the rich repertoire of functional patterns executed by motor neurons? How many neuronal populations arise from each cardinal domain and how many transcription factors are needed to univocally identify a cellular type? Attempts to answer these questions have resorted to lineage-tracing analysis to map the neurotransmitter phenotype, axonal projection patterns and functional roles during behavior of each of these different domains [Arber, 2000]; [Stepien & Arber, 2008]; [Goulding, 2009]. Neurons originating in the ventral domains are the most studied populations in the spinal cord. Among the ventral cardinal domains, insights into the mechanisms enacting neuronal diversification have arisen from the study of motor neurons [Jessell, 2000]. Motor neuron identity is linked to target muscle innervation. This relationship is so significant, that in 1925 Charles Sherrington proposed the concept of the motor unit -a motor neuron and the muscle fiber it innervates- the fundamental unit by which the nervous system controls movement [Kandel *et al*, 2012]. Groups of motor neurons forming selective connection with the same muscle belong to a "motor pool". Acquisition of genetic pool identity along the rostrocaudal axis is achieved through expression of Hox transcription factors at different spinal segments. A different set of Hox regulatory interactions directs motor pool diversity at a single segmental level [Dasen et al., 2005]; [Dasen & Jessell, 2009] (Figure 1.3c). Finally, the selection of target muscle connectivity is determined downstream of Hox regulation (Figure 1.3d).

The remaining progenitor domains give rise to interneuronal populations, containing an unknown number of cellular types, whose complete molecular identity continues imprecise. In what follows, we will review our current knowledge of interneuronal molecular diversity focusing on the identified four cardinal domains, V0 – V3. The V0 interneuronal population is derived from Dbx1 expressing progenitors, which settle in the ventromedial region of the mouse spinal cord and primarily project commissural axons. V0 interneurons can be divided into a ventral subpopulation which transiently express Evx1 (V0V) and a dorsal subpopulation (V0D) lacking expression of Evx1 [Pierani et al., 2001]; [Moran-Rivard et al., 2001] (Figure 1.4a). In addition, a small population of V0 interneurons has been identified, representing only a few percent of the parental population, delineated by the expression of the transcription factor Pitx2 which marks cholinergic cells that represents the sole source of C bouton inputs to motor neurons [Zagoraiou et al., 2009]. Tracing experiments have revealed further fractionation of the seemingly homogeneous V0c population, by discovering an ipsilaterally projecting population together with a bilaterally projecting population with motor neuron specific connectivity [Stepien et al., 2010]. These findings exemplify the importance of considering not only molecular phenotypes but also morphological features to fully characterize cellular types. The functional consequences of both populations remains to be explored.

V1 interneurons are defined by developmental expression of the homeodomain transcription factor En1 and, express GABA and/or glycine as inhibitory neurotransmitters [Saueressig et al., 1999]; [Sapir et al., 2004] (Figure 1.4b). This population constitutes over one third of all inhibitory interneurons in the ventral spinal cord, and include Renshaw cells and some Group Ia reciprocal interneurons (in this group, not all functionally defined neurons derive from a single progenitor domain), which mediate recurrent and reciprocal inhibition, respectively [Sapir et al, 2004] and [Zhang et al, 2014]. Renshaw interneurons the earliest populations to be defined physiologically [Renshaw, 1946] and for which transcription factors marking them have been isolated (Renshaw interneurons co-express Oc1, Oc2 and MafB, [Stam et al, 2012]). Yet these two physiologically defined subtypes represent only a small fraction of the parental V1 population [Alvarez et al, 2005], implying a greater diversity within V1 neurons yet unexplored. Moreover, molecular diversity remains unexplored even within functionally defined populations, a case best represented by the known morphological heterogeneity residing among Renshaw interneurons [Fyffe, 1990].

V2 interneurons are labelled initially by expression of Lhx3, and comprise ipsilaterally projecting excitatory V2a neurons [Crone et al., 2008], inhibitory V2b and V2c neurons [Panayi et al., 2010]. V2a and V2b cells are labelled by Chx10 and GATA3 respectively and are generated in equal numbers through a Notch-dependant mechanism [Okigawa et al, 2014], additional transcription factors regulates the balance between V2a-V2b types and contributes to diversification [Del Barrio et al., 2007]; [Joshi et al., 2009]; [Lee et al., 2008] (Figure 1.4c). V2c are labelled by Sox1. It is likely that additional diversification remains within these populations.

Lastly, V3 interneurons are a major group of excitatory commissural interneurons in the spinal cord, generated from the ventral-most progenitor domain that later migrates dorsally and laterally [Briscoe et al., 2000]; [Goulding et al, 2002]; [Zhang et al, 2008], [Carcagno et al, 2014] (Figure 1.4d). These neurons selectively express the transcription factor single-minded 1 (Sim1) upon becoming postmitotic [Zhang et al, 2008], [Blacklaws et al, 2015]. Cell types within V3 interneurons have not yet been characterized, however, in the mature mouse spinal cord Sim1-positive V3 INs gather into anatomically and electrophysiologically distinct populations located in the lower thoracic and upper lumbar regions [Borowska et al., 2013].

In conclusion, several studies have addressed the issue of interneuronal diversification for many of the ventral progenitor domains, but the extent of neuronal heterogeneity still remains to be fully explored. Even more, in a cardinal domains such as V1, more than eighty percent of its diversity remains unknown.

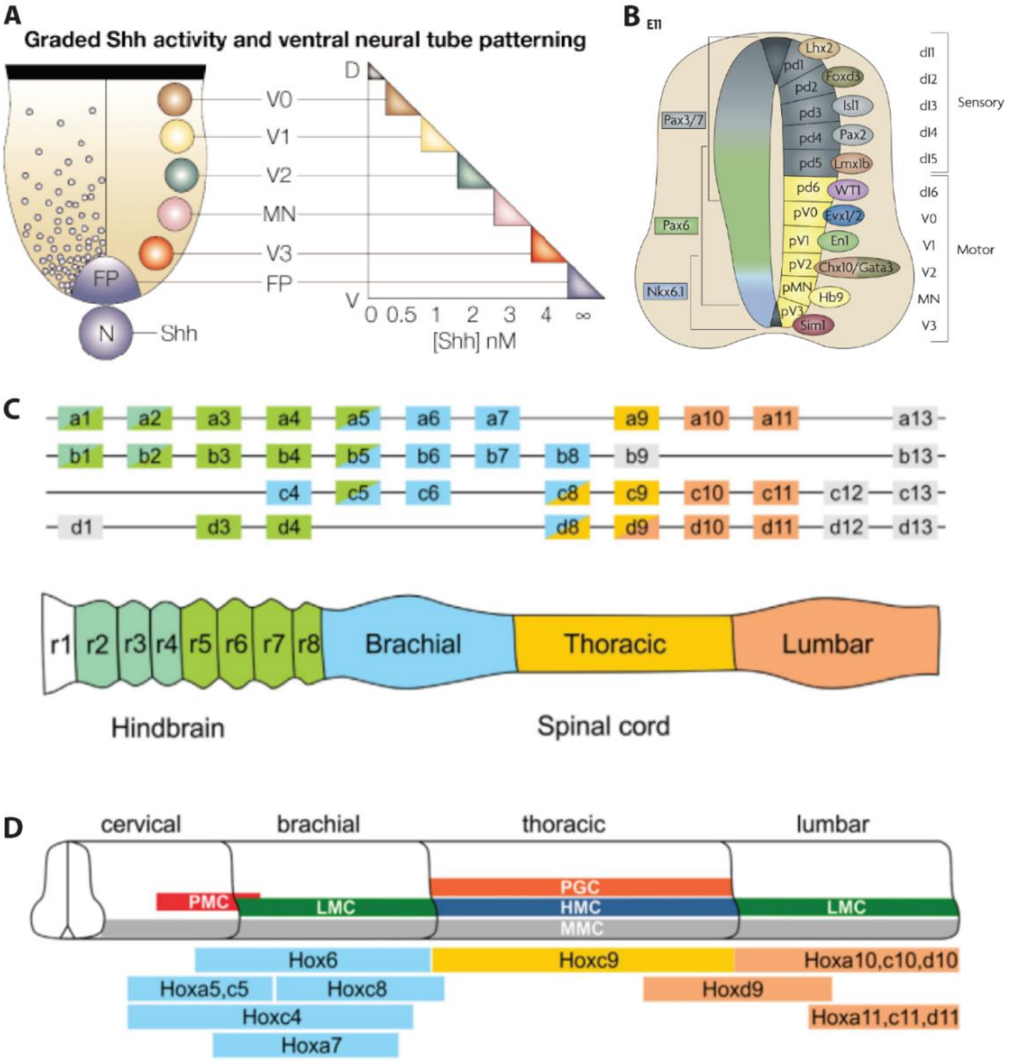


**Figure 1.3. The development of spinal progenitor domains and the specialization of motor pools.**
(a) Schematic representing the influence of Shh on the specification of ventral neuronal fates in the developing mouse spinal cord. The more dorsal the location of the progenitor domains in vivo, the lower the concentration of Shh, this in turn induces neuronal subpopulations in vitro (adapted from [Jessell, 2000]).
(b) Schematic cross-section of the mouse spinal cord at E11 showing eleven early classes of postmitotic neurons. dI1 - dI5 populations are originated from dorsal progenitors (sensory spinal pathways), whereas dI6, MN and V0–V3 develop from intermediate/ventral progenitors (locomotor pathways) (adapted from [Goulding, 2009])
(c) Expression Patterns of Hox genes in the hindbrain and spinal cord. Each Hox gene is expressed in discrete rostrocaudal domains within the hindbrain and spinal cord. Color coding of Hox genes represents expression domains along the rostrocaudal axis (adapted from [Philippidou and Dasen, 2013])

(d) Expression of Hox4–Hox11 genes in the spinal cord parallels motor neuron columnar and pool specification. Motor neurons in the lateral motor columns at brachial and lumbar levels further diversify into more than 50 pools innervating limb muscles (adapted from [Philippidou and Dasen, 2013])

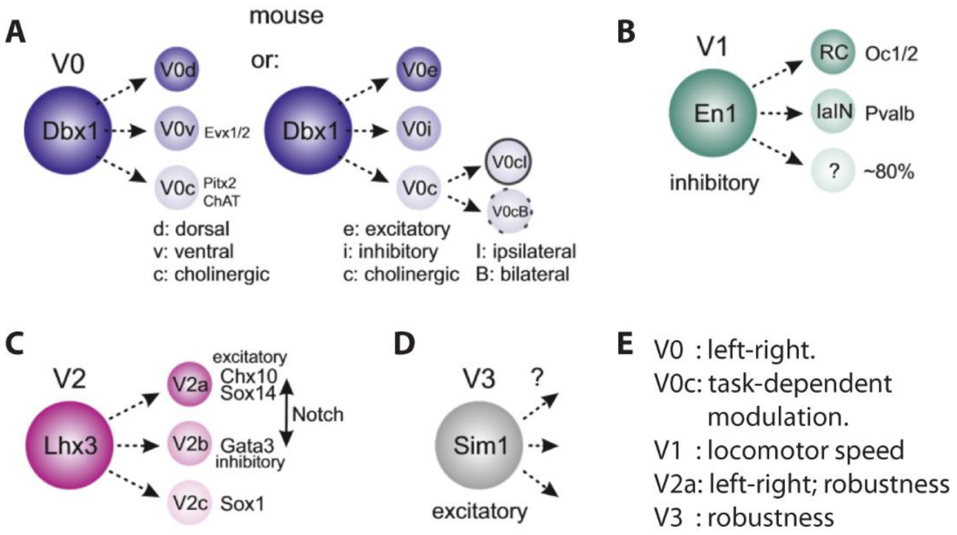


**Figure 1.4. Spinal interneuron diversity.**
Diversification of different interneuronal populations in the spinal cord (adapted from [Arber, 2012]).
(a) - (d) Populations contained within each cardinal domain (V0-V3).
(e) Function presumed to be performed by each population, resulting from genetic perturbation experiments.

*Functional correlates of ventral interneuronal diversity*

We have reviewed a series of genetic studies that aim to identify meaningful cellular types within each cardinal interneuronal population. However, such assessments must always link molecular phenotype with electrophysiological experiments to determine the functional correlates of each population. Alternatively, ablation or silencing studies could shine light on the significance of types in terms of alterations to behavior (see [Goulding, 2009]; [Kiehn, 2011]; [Stepien and Arber, 2008]). For example, inactivation of Dbx1 gene in mice was used to determine the role of V0 interneurons, establishing their necessity for proper coordination of left-right alternation during fictive locomotion [Lanuza et al., 2004]. Genetic inactivation of cholinergic V0c population suggests a modulatory role of motor neuron firing frequency and muscle activation and selective behavioral defects in motor performance during swimming but not locomotion [Zagoraiou et al., 2009]. Genetic ablation of V1 interneurons slows the speed of rhythmic locomotor output [Gosgnach et al., 2006], but in the absence of information on the diversity and connectivity of component V1 subtypes, the circuit logic that underlies such aberrant motor behavior has been difficult to resolve. Genetic ablation of V2a interneurons demonstrates that they serve an important function in the control of left-right alternation and are required to maintain robust locomotor patterns [Crone et al., 2008]. Finally, the less explored V3 interneuronal population is also needed to maintain a stable locomotor pattern [Zhang et al., 2008] (Figures 1.4e).

These experiments raise several open issues for future research. First, individual spinal progenitor domains are the source of many functionally diverse neuronal subpopulations. Perturbations therefore affect multiple descendant populations en bloc and may lead to defects that are difficult to interpret. More targeted genetic interference at the level of individual populations will be possible as soon as developmental maps are more closely aligned to subpopulation maps defined by electrophysiology and connectivity.

In summary, these studies have informed the genetic identity and functional relevance of distinct sets of interneurons that are necessary for controlling motor neuron activity during behavior. Nonetheless, individual cell type identification have proven hard. As a consequence, no connections have yet been established between a unique cell type and its behavioral function, a case best exemplified by perturbations performed on the entire V1 population known to contain Renshaw and Ia inhibitory interneurons;  it is difficult to predict how coincident elimination of both population affect behavior. It is worth noting that functional subtypes of interneurons appear to occupy stereotypic settling positions within the intermediate and ventral spinal cord [Thomas and Wilson, 1965]; [Hultborn et al., 1971]; [Hughes et al., 2005]. But it remains unclear whether the construction of interneuronal circuits takes advantage of neuronal settling position to establish local connectivity; a strategy used by motor neurons to form stereotypic patterns of proprioceptive input connectivity [Sürmeli et al., 2011]. Finally, it has yet to be resolved if the local circuits that control motor output adhere to a canonical wiring diagram, reiterated for each motor pool independent of the biomechanics of its limb muscle target.

## 1.5 Computational assessment of cellular diversity

To elucidate the organization of interneuronal populations in the spinal cord, an assessment of cellular heterogeneity has to be calculated. Estimates of cell type diversity can be obtained through a number of computational approaches. We provide here a brief review of these methods, organizing them into major research lines.

Computational methods employ clustering algorithms to arrange group of neurons into known classes or to identify the classes themselves, when measurements from features describing individual neurons are available (see [Armañanzas and Ascoli, 2015], for a recent review). These methods are effective, but have drawbacks. Current analysis based on clustering techniques are suboptimal, either because they are based on general data analysis methods, or because they do not adequately separate signal from noise. Relying on hierarchical clustering and dimensionality reduction methods (such as PCA or t-SNE; [Macosko et al, 2015]; [Jaitin et al, 2014]; [Grün and van Oudenaarden, 2015]), these general methods do not model the nuances of the data collection process and depend on fine-tuned parameters to analyze each individual dataset. More concretely, it is challenging to use hierarchical clustering methods to determine the number of cell types automatically because their inferences can be sensitive to the choice of similarity measures [Augen, 2005]. Even more critically, neither the account of uncertainty of estimates, nor the integration of different sources of information is involved in the aforementioned methods. In addition, a large number of cells need to be assessed to thoroughly sample and estimate the true extent of cell diversity within a population.

A different set of computational approaches based on deconvolution algorithms have been used to characterize cellular diversity from information about gene expression profiles [Shen-Orr and Gaujoux, 2013]. These can be divided into two major methodologies: Regression or matrix factorization approaches. Regression approaches can be applied when the expression profile of cell types of interest is known a priori [Wang et al, 2006]; [Abbas et al, 2009]; [Gong et al, 2011]; [Zuk et al, 2013]; [Grange, et al, 2014]. To overcome this limitation, a different breadth of regression algorithms relied on the premise that particular genes are expressed in only a single cell type [Gaujoux and Seoighe, 2012]. In cases in which cell types exhibit highly correlated transcriptional profiles, such optimization methods fail to resolve inherent heterogeneity. Different attempts to overcome this limitation by grouping candidate cell types with similar transcriptional profiles into classes can infer diversity at a class level, but lack the ability to identify individual cell types [Bullman et al, 2013]. The cases enumerated so far do not represent the general biological situation and are not applicable when the objective is de novo cell type discovery.

In contrast, matrix-factorization approaches become relevant when cell type expression profiles are not known [Repsilber, 2010]; [Erkkilä et al, 2010]; [Bazot et al, 2013]; [Zhong et al, 2013]; [Liebner et al, 2014]). However, these methodologies fail to adequately describe the underlying diversity, because under many conditions an infinite number of equally valid solutions exist for a particular dataset. This problem is accentuated when presumed cell types are present in a set of solutions but absent in a different one, making it impossible to assert any precise conclusion about their necessity to explain the data or their prevalence within the parental population.

Lastly, fueled by recent advances in sequencing technologies and RNA library preparations, clustering algorithms are used to distinguishing neuronal cell types from genome-wide approaches that assess mRNA expression profiles [Usoskin *et al*, 2015]; [Zeisel *et al*, 2015]; [Macosko *et al*, 2015]; [Tasic *et al*, 2016]. Nevertheless, documented dissociations between mRNA and protein expression [Gygi *et al*, 1999]; [Vogel & Marcotte, 2012] emphasize the merits of analysis of protein expression at the level of individual neurons [Sharma *et al*, 2015]. But if many genes are involved in defining individual subpopulations, then the validation of protein co-expression will be constrained by the limited repertoire of primary and secondary antibodies. As such, current computational approaches are not sufficient to tackle the problem of identifying interneuron diversity.

We will see in subsequent chapters that, in principle, the previously outlined practical limitations might be overcome with the development of statistical methods that are able to resolve the extent of neuronal diversity from sparsely sampled transcriptional datasets. We therefore sought out to develop statistical methods capable of revealing underlying diversity given incomplete data sampling - a common problem in biological systems, where complete descriptions of cellular characteristics are rarely available. Such a method might be expected to provide: (i) an objective measure of confidence in the existence of cell types and their prevalence within a parental population, (ii) improvement in estimation accuracy upon integrating independent cellular characteristics with molecular phenotype, and (iii) informative predictions that guide further experiments to improve estimates of cellular diversity.

# Chapter 2

# Molecular characterization of V1 spinal interneurons

## 2.1 Introduction

V1 interneurons represent almost a third of all inhibitory interneurons within the ventral spinal cord. They are necessary for the regulation of the locomotor step cycle and, the shaping of motor outputs during locomotion [Gosgnach et al, 2006]. While this class is known to contain two physiologically defined interneuron subtypes, Renshaw and group Ia interneurons that mediate recurrent and reciprocal inhibition of motor neurons, respectively, these constitute only a small fraction of all V1 interneurons. The total number of cell types within V1 and the exact contribution of individual V1 interneuron cell types to regulating locomotion remains to be determined.

In this chapter I describe the molecular characterization of V1 interneurons by their transcription factor profile. By using immunohistochemical measurements, we assess the prevalence of each factor within the parental population, as well as pairwise coexpression. To assess the extent of diversity within this population, we resort to a statistical framework capable of inferring the underlying cell types by assigning the most likely expression profiles consistent with the data while assigning confidence on their existence. Finally, we show that just pairwise transcriptional information is not enough to delineate the full cell type repertoire with high enough confidence; an issue that we will address in subsequent chapters.

# 2.2 Results

## 2.2.1 Transcriptional diversity of V1 inhibitory interneurons

To explore the diversity of spinal inhibitory interneurons we compared the gene expression profiles of En1$^+$ V1 and Ptf1a$^+$ dI4/dILA interneurons [Saueressig et al., 1999]; [Sapir et al., 2004]; [Glasgow et al., 2005] (Figure 1.6a). V1 interneurons provide prominent synaptic input to motor neurons, whereas dI4/dILA interneurons shun motor neurons and instead select sensory terminals as ventral targets [Betley et al., 2009]; [Goulding et al., 2014]. We reasoned that a comparison of these two inhibitory populations, each settling in a different dorsoventral position and forming distinct postsynaptic targets, should reveal genes that fractionate the parental V1 and dI4/dILA populations, while excluding generic markers expressed by both inhibitory populations. We focused this analysis on transcription factors (TFs), given their roles in specifying neuronal subtype identity and connectivity [Dalla Torre di Sanguinetto et al., 2008]; [Amamoto and Arlotta, 2014].

The V1 and dI4/dILA interneuron sets were isolated by fluorescence activated cell sorting, from spinal cords dissociated from *En1::cre; Rosa.lsl.eYFP* and *Ptf1a::cre; Rosa.lsl.eYFP* mice respectively. To accommodate the possibility of dynamic changes in gene expression, microarrays were performed at three different ages - e12.5, p0, and p5 - a developmental window that covers the emergence of interneuron identity and the formation of synaptic connections (Figures 1.6b, c). Comparative microarray analysis identified 56 genes that encoded TFs with a >3-fold enrichment in V1 interneurons (mean V1:dI4 enrichment: 74.5; range: 3.1 to 930-fold; $p \leq 0.02$ by one-way ANOVA) at one or more developmental stages, and 160 TF genes with a >3-fold enrichment in dI4/dILA interneurons (mean dI4:V1 enrichment: 38.5; range: 3.1 to 784-fold; $p \leq 0.02$ by one-way ANOVA). We focus on diversity within the parental V1 interneuron population.

Analysis of gene expression databases [Visel et al., 2004]; [Sunkin et al., 2013] revealed that 32 of the 56 V1 TFs exhibited mosaic expression in the embryonic (e11.5-e15.5) or neonatal (p4) ventral spinal cord. Two additional genes, MafA and Prox1, exhibit scattered expression in the ventral spinal cord [Misra et al., 2008]; [Lecoin et al., 2010], and were therefore included in subsequent analyses. From these 34 candidates, we focused on 19 TFs for which we were able to generate (FoxP1, FoxP2, FoxP4, Lmo3, MafA, Nr3b2, Nr4a2, Nr5a2, Otp, Pou6f2, Prdm8, and Sp8) or obtain (bhlhb5, MafB, Nr3b3, Oc1, Oc2, Prox1, and Zfhx4) antibodies that permit immunohistochemical.

Where possible, antibody specificity was validated by showing an absence of staining in knockout animals. Antibodies previously confirmed to be specific against knockout mice include: anti-Bhlhb5 [Ross et al., 2010], antiFoxP1 [Sürmeli et al., 2011], anti-Nr3b2 [Chen and Nathans, 2007], anti-Onecut1 [Wu et al., 2012], and anti-Prdm8 [Ross et al., 2012]. Additionally, all FoxP2, Sp8, Otp, and Pou6f2 antibodies used in this study were validated against knockout mice (Figure 1.6d-g and data not shown). The Sigma rabbit anti-MafB antibody recognizes MafB, but may also weakly detect other Maf-family members (F.J.Alvarez, unpublished observation), as immunoreactivity is not completely abolished in MafB::GFP homozygous null mice. In cases where knockout mice were unavailable, we confirmed that the pattern of immunoreactivity was consistent with previously published expression data obtained from the Allen Brain Institute

(Figure 1.6h-o), Website: ©2012 Allen Institute for Brain Science. Allen Spinal Cord Atlas. Available from: http://mousespinal.brain-map.org/.

To evaluate the prevalence of these 19 TFs within the parental V1 population we marked V1 interneurons by LacZ expression in *En1::cre; Tau.lsl.nlacZ* (En1.nLacZ) mice, and performed dual immunohistochemical analysis with antibodies directed against LacZ and each TF individually. We focused our analysis on p0 lumbar spinal cord, a time after the specification and migration of interneurons is complete but before extinction of expression of many developmentally regulated TFs. We found that all 19 TFs were expressed in subsets of V1 interneurons, with the incidence of expression ranging from 5% (MafA) to 74% (Lmo3) of the parental V1 population. Conjoint exposure to antibodies directed against 14 of these TFs marked 90% of V1 interneurons in p0 lumbar spinal segments, indicative of near-complete coverage of the parental V1 population. In addition, we detected expression of 12 of the 19 V1 TFs within subsets of V2a excitatory interneurons (threshold: 3%; range: 3-62% of parental V2a interneurons), indicating that this set of TFs reveal diversity in both inhibitory and excitatory interneurons (Figure 1.7a).

Next, the extent of co-expression was determined for various binary TF combinations, by performing triple antibody staining, with each pairing gated to LacZ+ neurons in p0 En1.LacZ mice (Figure 1.7b-c). Out of 171 possible TF combinations, we were able to measure 148; the complete analysis was hindered due to antibody incompatibility, in which primary antibodies generated in the same host species cannot be distinguished easily by fluorescently-tagged secondary antibodies (host co-reactivity). This analysis identified four TFs - FoxP2, MafA, Pou6f2, and Sp8 - delineating non-overlapping sets that cumulatively comprise 64% of the parental V1 population. Given this single and pairwise TF information, we then resort to a statistical formalism to estimate the entire profile of possible cell types consistent with this data.

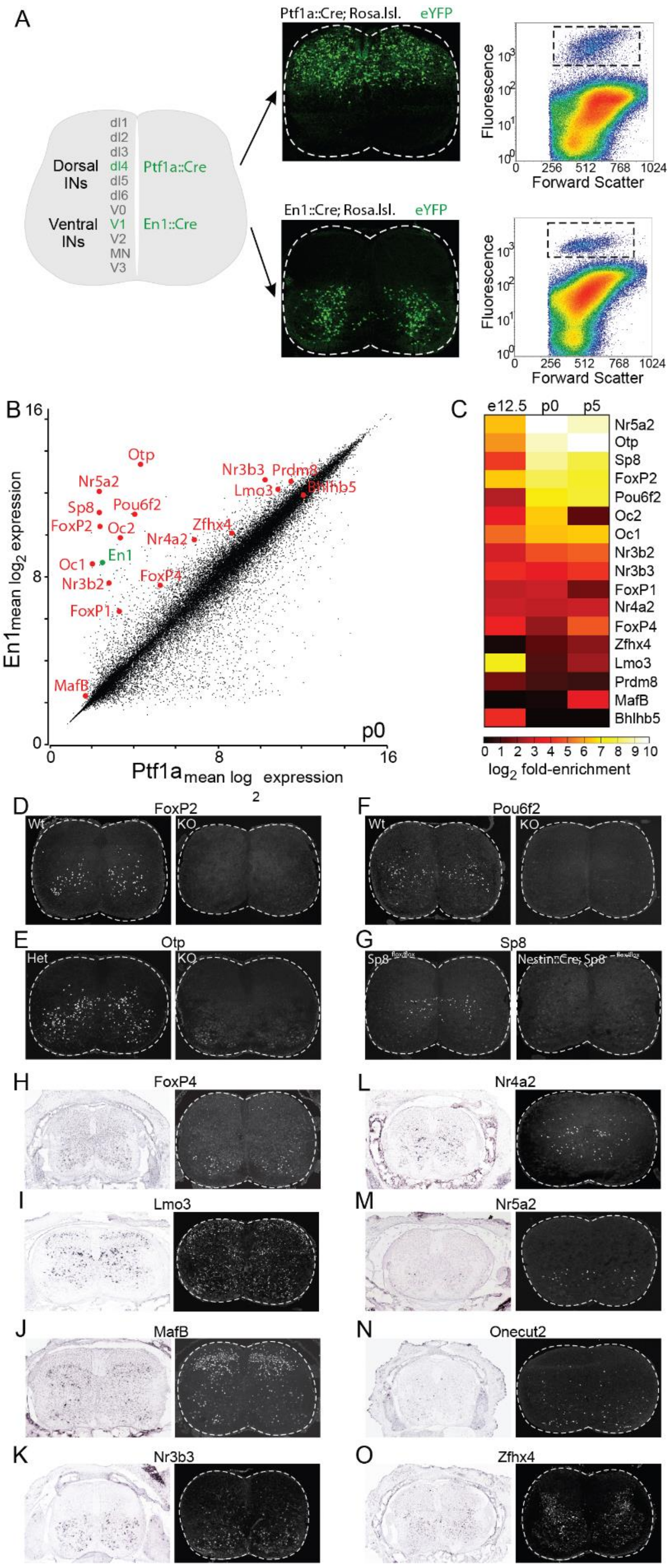


**Figure 2.1. Transcription Factors Enriched in V1 Interneurons and Characterization of Antibody Specificity.** (a) Isolation of V1 and dI4/dILA interneurons. Left, interneuron populations. Middle, En1::Cre (V1) and Ptf1a::Cre (dI4/dILA) lineage-traced interneurons in p0 lumbar spinal cord. Right, FACS-isolated eYFP+ interneurons for microarray analysis.

(b) Scatter plot of expression levels of transcription factors (TFs, red) enriched in V1 interneurons from p0 mice.

(c) TFs with $>$ 3-fold enrichment ($p \leq 0.02$, one-way ANOVA) at one or more developmental ages.

(d)- (g) Immunohistochemical validation of antibody specificity, demonstrating absence of immunoreactivity in knockout mice.

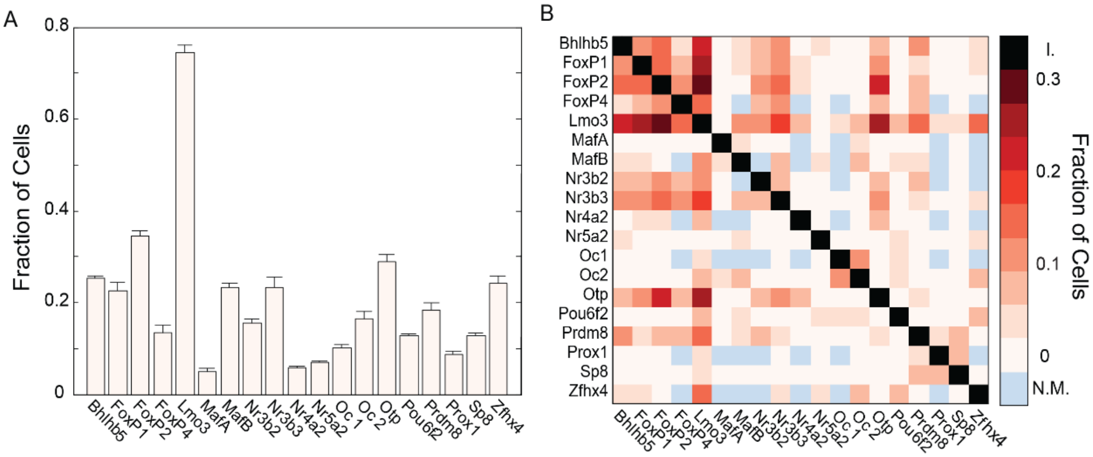
(d) Goat anti-FoxP2, p6 wild-type or FoxP2::Flpo KO mice.

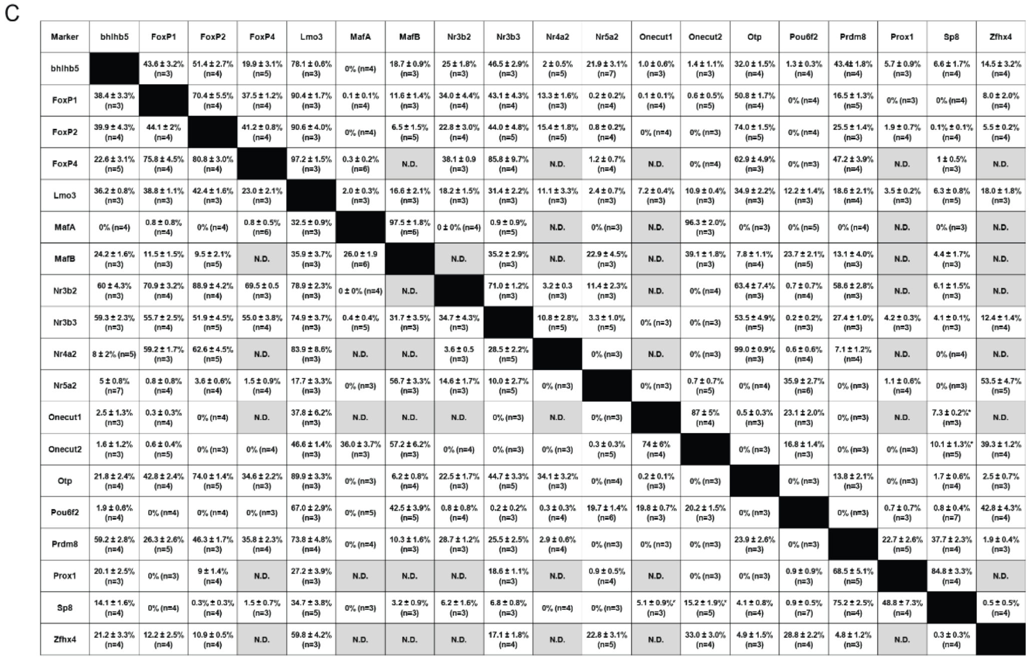
(e) Guinea pig anti-Otp, e17.5 Otp::Flpo heterozygous (control) or homozygous (knockout) mice.

(f) Rabbit anti-Pou6f2, p3 wild-type or Pou6f2 knockout mice.

(g) Goat anti-Sp8, p4 Sp8flox/flox or Nestin::Cre; Sp8flox/flox mice. Similar specificity was seen with FoxP2, Otp, Pou6f2, and Sp8 antibodies generated in other species. All images represent lumbar (L3-L5) spinal segments.

(h)- (o) Left, representative examples of expression in p4 mouse spinal cord by in situ hybridization, (Data Reproduced from Allen institute for Brain Science. Right, representative examples of expression in spinal cord by immunohistochemistry.





| Marker | bhlhb5 | FoxP1 | FoxP2 | FoxP4 | Lmo3 | MafA | MafB | Nr3b2 | Nr3b3 | Nr4a2 | Nr5a2 | Onecut1 | Onecut2 | Otp | Pou6f2 | Prdm8 | Prox1 | Sp8 | Zfhx4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| bhlhb5 |  | 43.6 ± 3.2%<br>(n=3) | 51.4 ± 2.7%<br>(n=4) | 19.9 ± 3.1%<br>(n=5) | 78.1 ± 0.6%<br>(n=3) | 0% (n=4) | 18.7 ± 0.9%<br>(n=3) | 25 ± 1.8%<br>(n=3) | 46.5 ± 2.9%<br>(n=3) | 2 ± 0.5%<br>(n=5) | 21.9 ± 3.1%<br>(n=7) | 1.0 ± 0.6%<br>(n=3) | 1.4 ± 1.1%<br>(n=3) | 32.0 ± 1.5%<br>(n=4) | 1.3 ± 0.3%<br>(n=4) | 43.4± 1.8%<br>(n=4) | 5.7 ± 0.9%<br>(n=3) | 6.6 ± 1.7%<br>(n=4) | 14.5 ± 3.2%<br>(n=4) |
| FoxP1 | 38.4 ± 3.3%<br>(n=3) |  | 70.4 ± 5.5%<br>(n=4) | 37.5 ± 1.2%<br>(n=4) | 90.4 ± 1.7%<br>(n=3) | 0.1 ± 0.1%<br>(n=4) | 11.6 ± 1.4%<br>(n=3) | 34.0 ± 4.4%<br>(n=4) | 43.1 ± 4.3%<br>(n=4) | 13.3 ± 1.6%<br>(n=3) | 0.2 ± 0.2%<br>(n=4) | 0.1 ± 0.1%<br>(n=4) | 0.6 ± 0.5%<br>(n=5) | 50.8 ± 1.7%<br>(n=4) | 0% (n=4) | 16.5 ± 1.3%<br>(n=5) | 0% (n=3) | 0% (n=4) | 8.0 ± 2.0%<br>(n=4) |
| FoxP2 | 39.9 ± 4.3%<br>(n=4) | 44.1 ± 2%<br>(n=4) |  | 41.2 ± 0.8%<br>(n=4) | 90.6 ± 4.0%<br>(n=3) | 0% (n=4) | 6.5 ± 1.5%<br>(n=5) | 22.8 ± 3.0%<br>(n=4) | 44.0 ± 4.8%<br>(n=5) | 15.4 ± 1.8%<br>(n=5) | 0.8 ± 0.2%<br>(n=4) | 0% (n=4) | 0% (n=3) | 74.0 ± 1.5%<br>(n=5) | 0% (n=4) | 25.5 ± 1.4%<br>(n=3) | 1.9 ± 0.7%<br>(n=4) | 0.1% ± 0.1%<br>(n=4) | 5.5 ± 0.2%<br>(n=4) |
| FoxP4 | 22.6 ± 3.1%<br>(n=5) | 75.8 ± 4.5%<br>(n=4) | 80.8 ± 3.0%<br>(n=4) |  | 97.2 ± 1.5%<br>(n=3) | 0.3 ± 0.2%<br>(n=6) | N.D. | 38.1 ± 0.9<br>(n=3) | 85.8 ± 9.7%<br>(n=4) | N.D. | 1.2 ± 0.7%<br>(n=4) | N.D. | 0% (n=4) | 62.9 ± 4.9%<br>(n=3) | 0% (n=3) | 47.2 ± 3.9%<br>(n=4) | N.D. | 1 ± 0.5%<br>(n=3) | N.D. |
| Lmo3 | 36.2 ± 0.8%<br>(n=3) | 38.8 ± 1.1%<br>(n=3) | 42.4 ± 1.6%<br>(n=3) | 23.0 ± 2.1%<br>(n=3) |  | 2.0 ± 0.3%<br>(n=3) | 16.6 ± 2.1%<br>(n=3) | 18.2 ± 1.5%<br>(n=3) | 31.4 ± 2.2%<br>(n=3) | 11.1 ± 3.3%<br>(n=3) | 2.4 ± 0.7%<br>(n=3) | 7.2 ± 0.4%<br>(n=3) | 10.9 ± 0.4%<br>(n=3) | 34.9 ± 2.2%<br>(n=3) | 12.2 ± 1.4%<br>(n=3) | 18.6 ± 2.1%<br>(n=4) | 3.5 ± 0.2%<br>(n=3) | 6.3 ± 0.8%<br>(n=5) | 18.0 ± 1.8%<br>(n=3) |
| MafA | 0% (n=4) | 0.8 ± 0.8%<br>(n=4) | 0% (n=4) | 0.8 ± 0.5%<br>(n=6) | 32.5 ± 0.9%<br>(n=3) |  | 97.5 ± 1.8%<br>(n=6) | 0 ± 0% (n=4) | 0.9 ± 0.9%<br>(n=5) | N.D. | 0% (n=3) | N.D. | 96.3 ± 2.0%<br>(n=3) | 0% (n=3) | 0% (n=5) | 0% (n=4) | N.D. | 0% (n=3) | N.D. |
| MafB | 24.2 ± 1.6%<br>(n=3) | 11.5 ± 1.5%<br>(n=3) | 9.5 ± 2.1%<br>(n=5) | N.D. | 35.9 ± 3.7%<br>(n=3) | 26.0 ± 1.9<br>(n=6) |  | N.D. | 35.2 ± 2.9%<br>(n=3) | N.D. | 22.9 ± 4.5%<br>(n=3) | N.D. | 39.1 ± 1.8%<br>(n=3) | 7.8 ± 1.1%<br>(n=4) | 23.7 ± 2.1%<br>(n=5) | 13.1 ± 4.0%<br>(n=3) | N.D. | 4.4 ± 1.7%<br>(n=3) | N.D. |
| Nr3b2 | 60 ± 4.3%<br>(n=3) | 70.9 ± 3.2%<br>(n=4) | 88.9 ± 4.2%<br>(n=4) | 69.5 ± 0.5<br>(n=3) | 78.9 ± 2.3%<br>(n=3) | 0 ± 0% (n=4) | N.D. |  | 71.0 ± 1.2%<br>(n=3) | 3.2 ± 0.3<br>(n=3) | 11.4 ± 2.3%<br>(n=3) | N.D. | 0% (n=4) | 63.4 ± 7.4%<br>(n=3) | 0.7 ± 0.7%<br>(n=4) | 58.6 ± 2.8%<br>(n=3) | N.D. | 6.1 ± 1.5%<br>(n=3) | N.D. |
| Nr3b3 | 59.3 ± 2.3%<br>(n=3) | 55.7 ± 2.5%<br>(n=4) | 51.9 ± 4.5%<br>(n=5) | 55.0 ± 3.8%<br>(n=4) | 74.9 ± 3.7%<br>(n=3) | 0.4 ± 0.4%<br>(n=5) | 31.7 ± 3.5%<br>(n=3) | 34.7 ± 4.3%<br>(n=3) |  | 10.8 ± 2.8%<br>(n=5) | 3.3 ± 1.0%<br>(n=5) | 0% (n=3) | 0% (n=3) | 53.5 ± 4.9%<br>(n=5) | 0.2 ± 0.2%<br>(n=3) | 27.4 ± 1.0%<br>(n=3) | 4.2 ± 0.3%<br>(n=3) | 4.1 ± 0.1%<br>(n=3) | 12.4 ± 1.4%<br>(n=4) |
| Nr4a2 | 8 ± 2% (n=5) | 59.2 ± 1.7%<br>(n=3) | 62.6 ± 4.5%<br>(n=5) | N.D. | 83.9 ± 8.6%<br>(n=3) | N.D. | N.D. | 3.6 ± 0.5<br>(n=3) | 28.5 ± 2.2%<br>(n=5) |  | 0% (n=3) | N.D. | 0% (n=3) | 99.0 ± 0.9%<br>(n=3) | 0.6 ± 0.6%<br>(n=4) | 7.1 ± 1.2%<br>(n=4) | N.D. | 0% (n=4) | N.D. |
| Nr5a2 | 5 ± 0.8%<br>(n=7) | 0.8 ± 0.8%<br>(n=4) | 3.6 ± 0.6%<br>(n=4) | 1.5 ± 0.9%<br>(n=4) | 17.7 ± 3.3%<br>(n=3) | 0% (n=3) | 56.7 ± 3.3%<br>(n=3) | 14.6 ± 1.7%<br>(n=3) | 10.0 ± 2.7%<br>(n=5) | 0% (n=3) |  | 0% (n=3) | 0.7 ± 0.7%<br>(n=5) | 0% (n=4) | 35.9 ± 2.7%<br>(n=6) | 0% (n=3) | 1.1 ± 0.6%<br>(n=4) | 0% (n=3) | 53.5 ± 4.7%<br>(n=5) |
| Onecut1 | 2.5 ± 1.3%<br>(n=3) | 0.3 ± 0.3%<br>(n=4) | 0% (n=4) | N.D. | 37.8 ± 6.2%<br>(n=3) | N.D. | N.D. | N.D. | 0% (n=3) | N.D. | 0% (n=3) |  | 87 ± 5%<br>(n=4) | 0.5 ± 0.3%<br>(n=3) | 23.1 ± 2.0%<br>(n=3) | 0% (n=3) | N.D. | 7.3 ± 0.2%*<br>(n=3) | N.D. |
| Onecut2 | 1.6 ± 1.2%<br>(n=3) | 0.6 ± 0.4%<br>(n=5) | 0% (n=3) | 0% (n=4) | 46.6 ± 1.4%<br>(n=3) | 36.0 ± 3.7%<br>(n=3) | 57.2 ± 6.2%<br>(n=3) | 0% (n=4) | 0% (n=3) | 0% (n=3) | 0.3 ± 0.3%<br>(n=5) | 74 ± 6%<br>(n=4) |  | 0% (n=3) | 16.8 ± 1.4%<br>(n=3) | 0% (n=3) | 0% (n=3) | 10.1 ± 1.3%*<br>(n=5) | 39.3 ± 1.2%<br>(n=4) |
| Otp | 21.8 ± 2.4%<br>(n=4) | 42.8 ± 2.4%<br>(n=4) | 74.0 ± 1.4%<br>(n=5) | 34.6 ± 2.2%<br>(n=3) | 89.9 ± 3.3%<br>(n=3) | 0% (n=3) | 6.2 ± 0.8%<br>(n=4) | 22.5 ± 1.7%<br>(n=3) | 44.7 ± 3.3%<br>(n=5) | 34.1 ± 3.2%<br>(n=3) | 0% (n=4) | 0.2 ± 0.1%<br>(n=3) | 0% (n=3) |  | 0% (n=3) | 13.8 ± 2.1%<br>(n=3) | 0% (n=3) | 1.7 ± 0.6%<br>(n=4) | 2.5 ± 0.7%<br>(n=3) |
| Pou6f2 | 1.9 ± 0.6%<br>(n=4) | 0% (n=4) | 0% (n=4) | 0% (n=3) | 67.0 ± 2.9%<br>(n=3) | 0% (n=5) | 42.5 ± 3.9%<br>(n=5) | 0.8 ± 0.8%<br>(n=4) | 0.2 ± 0.2%<br>(n=3) | 0.3 ± 0.3%<br>(n=4) | 19.7 ± 1.4%<br>(n=6) | 19.8 ± 0.7%<br>(n=3) | 20.2 ± 1.5%<br>(n=3) | 0% (n=3) |  | 0% (n=3) | 0.7 ± 0.7%<br>(n=3) | 0.8 ± 0.4%<br>(n=7) | 42.8 ± 4.3%<br>(n=4) |
| Prdm8 | 59.2 ± 2.8%<br>(n=4) | 26.3 ± 2.6%<br>(n=5) | 46.3 ± 1.7%<br>(n=3) | 35.8 ± 2.3%<br>(n=4) | 73.8 ± 4.8%<br>(n=4) | 0% (n=4) | 10.3 ± 1.6%<br>(n=3) | 28.7 ± 1.2%<br>(n=3) | 25.5 ± 2.5%<br>(n=3) | 2.9 ± 0.6%<br>(n=4) | 0% (n=3) | 0% (n=3) | 0% (n=3) | 23.9 ± 2.6%<br>(n=3) | 0% (n=3) |  | 22.7 ± 2.6%<br>(n=5) | 37.7 ± 2.3%<br>(n=4) | 1.9 ± 0.4%<br>(n=3) |
| Prox1 | 20.1 ± 2.5%<br>(n=3) | 0% (n=3) | 9 ± 1.4%<br>(n=4) | N.D. | 27.2 ± 3.9%<br>(n=3) | N.D. | N.D. | N.D. | 18.6 ± 1.1%<br>(n=3) | N.D. | 0.9 ± 0.5%<br>(n=4) | N.D. | 0% (n=3) | 0% (n=3) | 0.9 ± 0.9%<br>(n=3) | 68.5 ± 5.1%<br>(n=5) |  | 84.8 ± 3.3%<br>(n=4) | N.D. |
| Sp8 | 14.1 ± 1.6%<br>(n=4) | 0% (n=4) | 0.3% ± 0.3%<br>(n=4) | 1.5 ± 0.7%<br>(n=3) | 34.7 ± 3.8%<br>(n=5) | 0% (n=3) | 3.2 ± 0.9%<br>(n=3) | 6.2 ± 1.6%<br>(n=3) | 6.8 ± 0.8%<br>(n=3) | 0% (n=4) | 0% (n=3) | 5.1 ± 0.9%*<br>(n=3) | 15.2 ± 1.9%*<br>(n=5) | 4.1 ± 0.8%<br>(n=4) | 0.9 ± 0.5%<br>(n=7) | 75.2 ± 2.5%<br>(n=4) | 48.8 ± 7.3%<br>(n=4) |  | 0.5 ± 0.5%<br>(n=4) |
| Zfhx4 | 21.2 ± 3.3%<br>(n=4) | 12.2 ± 2.5%<br>(n=4) | 10.9 ± 0.5%<br>(n=4) | N.D. | 59.8 ± 4.2%<br>(n=3) | N.D. | N.D. | N.D. | 17.1 ± 1.8%<br>(n=4) | N.D. | 22.8 ± 3.1%<br>(n=5) | N.D. | 33.0 ± 3.0%<br>(n=4) | 4.9 ± 1.5%<br>(n=3) | 28.8 ± 2.2%<br>(n=4) | 4.8 ± 1.2%<br>(n=3) | N.D. | 0.3 ± 0.3%<br>(n=4) |  |

**Figure 2.2. Prevalence of Identified Transcription Factors within V1 Interneurons.**
(a) Measured fraction of En1+ neurons labeled by each of the 19 individual transcription factors, in p0 lumbar spinal cord. Mean ± SEM, n ≥ 3 animals.

(b) Measured fraction of En1+ neurons labeled by pairs of transcription factors. Some comparisons were not possible due to antibody incompatibility (N.M., Not Measured). Diagonal values represent identity. Mean ± SEM, n ≥ 3 animals.

(c) Measured fraction of En1+ neurons labeled by pairs of transcription factors depicted on table format with exact values. Again, N.D. indicates values not measured due to antibody incompatibility. Diagonal values represent identity. Mean ± SEM, n ≥ 3 animals.

### 2.2.2 A sparse Bayesian approach for uncovering neuronal diversity

We developed a statistical analysis of V1 interneuron heterogeneity informed, initially, by two sets of data: (i) the fraction of neurons within the V1 parental population that express each of the 19 transcription factors, (ii) the fractions of neurons co-expressing various pairs of these transcription factors (We will see in subsequent chapters that positional distribution of V1 interneurons expressing each of the transcription factors was highly informative to constrain our inference procedure). The goal of this analysis is to determine the number of cell types needed to explain the expression and co-expression data. It is not yet tractable experimentally to delineate all higher-than-pairwise transcription factor combinations, given the vast number of potential combinations and limitations in antibody and fluorophore repertoire. We therefore developed an approach that permits statistical inference on the basis of partial analysis of the parental V1 population. Such inferences can be used to indicate which subsets of triple and higher-order genetic combination are most informative for further experimental study. This last feature has proven useful even at the pairwise level, since complete cytochemical analysis of all transcription factor pairs is hindered by antibody incompatibility.

In this statistical analysis, a cell type is defined by the pattern of expression of the 19 transcription factors under consideration. We characterize transcription factors as either expressed or not expressed, and thus each expression pattern, for example pattern k, is specified by a vector of 19 binary numbers, $J_{k,a}$, with $a$ ranging from 1 to 19. $J_{k,a}$ is set to 1 if transcription factor a is expressed as part of expression pattern k, and to 0 if it is not. This results in $2^{19}$ possible binary expression patterns for 19 transcription factors. This large number can be reduced by eliminating combinations that include pairs of factors never observed to be co-expressed within the same neuron. Our analysis of the pairwise expression data revealed that 67 out of 148 measured transcription factor pairs fail to co-express (Figure 2.2b-c). In fact, the criterion used to identify absence of co-expression was to consider any pairwise fractional value below one percent to zero. This information reduces the potential diversity to 1,978 potential expression patterns. Thus, for the variables $J_{k,a}$ specifying expression patterns, k runs from 1 to 1,978.

We assigned expression patterns to these cell types under the assumption that the total number is far fewer than the 1,978 potential expression patterns (and thus potential cell types). To achieve this we introduce cell-type fractions, $f_k$, with k again ranging across all the potential expression patterns (1 to 1,978): $f_k$ is the fraction of V1 interneurons with expression pattern k (equivalently, the fraction of neurons of cell-type k). Cell-type fractions must be positive ($f_k \geq 0$) and sum to 1 ($\Sigma_\mathrm{k}\ f_k = 1$), indicating that the entire V1 population is accounted for. The fraction of V1 neurons expressing transcription factor *a* (the data in Figure 2.2a) is $\Sigma_\mathrm{k}\ f_k\ J_{k,a}$, and the fraction co-expressing factors a and b (data in Figure 2.2b) is $\Sigma_\mathrm{k}\ f_k\ J_{k,a}\ J_{k,b}$.

*The cell type not expressing any gene*

Since the cell type not expressing any transcription factor has all its $J_{k,a}$ coefficients equal to zero, its presence in the previous sums would have no contribution and can therefore be excluded. Nonetheless, its fraction in the population can be computed indirectly by subtracting from 1 the sum of all the other cell types $f_{1978} = 1 - \Sigma_\mathrm{k}\ f_k$.

*Mathematical formalism*

Mathematically, we represent the mean measured values for each transcription factor as $M^1_i$ and the standard deviation as $\sigma_a$, and according with our previous definitions, $M^1_a$ is a linear combination of the unobserved populations of cells expressing all the possible binary vectors $J$ plus Gaussian noise. Similarly, measurements on the fraction of neurons co-expressing pairs of transcription factors are denoted by $M^2_{a, b}$ and $\sigma_{a,b}$, representing means and standard deviations.

$$M^1_a = \Sigma_k f_k J_{k,a} + \varepsilon_a \qquad \varepsilon_a \sim N(0,\sigma^2_a) \qquad a = 1,...,19$$

$$M^2_{a,b} = \Sigma_k f_k J_{k,a} J_{k,b} + \varepsilon_{ab} \qquad \varepsilon_{ab} \sim N(0,\sigma^2_{ab}) \qquad a, b = 1,...,19 \tag{2.1}$$

Subject to:

$$0 \le M^1_a \le 1 \;,\; 0 \le M^2_{ab} \le 1 \;,\; f_k \ge 0 \;,\; \Sigma_k f_k \le 1 \tag{2.2}$$

Fitting data within this framework amounts to choosing a set of cell-type fractions that provide a good match to the expression and co-expression data and that satisfy the positive / sum-to-one constraints. How can the results of such a fit best be evaluated? The number of non-zero inferred cell type fractions determines the predicted number of cell types, and the variables $J_{k,a}$ for a = 1, ... 19 and for k values with $f_k \neq 0$, determine the expression patterns of these selected cell types. In reality, however, the interpretation of this model turns out to be more nuanced. Before searching for a resolution for the inferred cell types, we rewrite the equations in (2.1) in a more compact representation by dividing by the noise standard deviation and grouping the non-zero measurements as a standardized linear regression problem.

$$Y_i = \Sigma_k q_{k,i} f_k + \varepsilon_i \quad ; \quad \varepsilon_i \sim N(0, \sigma^2_i) \quad ; \quad i = 1,...,148 \;,\; k =1,\ldots,1978 \tag{2.3}$$

Subject to:

$$0 \leq Y_i \leq 1 \ , \quad f_k \geq 0 \ , \quad \Sigma_k f_k = 1 \quad \text{and} \quad \Sigma_k q_{k,A} f_k \geq 0 \tag{2.4}$$

in which the vector Y represents our measured data $M^1_a$ and $M^2_{a,b}$ normalized by the corresponding $\sigma_a$ and $\sigma_{ab}$, and $q_{k,i}$ are the normalized coefficients $J_{k,a}$ or $J_{k,a} J_{k,b}$. Additionally, we impose the last constraint in (2.4), in contrast to (2.2), for all indices A for which the corresponding $M^1_a$ and $M^2_{ab}$ are different from zero, requiring non-zero mean predicted values for non-zero measured populations.

*Non-Negative Constrained Least Squares*

In principle, the model could be fit to the observed data by minimizing the summed squared difference between the measurements and the predictions generated by the inferred fractions. This methodology amounts to a non-negative constrained least squares (NNCLS) minimization problem [Wang et al, 2006]; [Abbas et al, 2009]; [Gong et al, 2011]; [Grange, et al, 2014].

$$f^* = \text{argmin}_f \ ||Y - Qf||^2_2 \tag{2.5}$$

Subject to:

$$f_k \geq 0 \ , \quad \Sigma_k f_k \leq 1 \quad \text{and} \quad \Sigma_k q_{k,A} f_k \geq Y_i \text{-}3 \tag{2.6}$$

Where $Y$, $f$ and $Q$ are vectors and a matrix with components $Y_i$, $f_k$ and $q_{i,k}$. The last constraint in (2.6) replaces (2.4) in our implementation, to comply with the inequality when using a numerical convex solver; we constrain the inferred measured values to be within three standard deviations of their measured mean, which is reasonably conservative given the Gaussian noise model. For approximate solutions to this type of optimization problems, which scale well when the dimension K of the vector *f* is large, see [Zuk et al., 2013]. In our case, K=1978, and the problem was still amenable to exact solutions. As in the LASSO, the first two constraints in (2.6) impose sparsity [Tibshirani, 1996]; [Meinshausen et al., 2013].

The NNCLS approach fails to describe adequately the extent of V1 interneuron diversity, because despite the constraint of nonnegativity and the sparsity imposition, an infinite number of equally valid solutions exist for our transcriptional data. Indeed, for any single presumed cell type it is possible to find alternative solutions that exclude this cell type while maintaining an optimal summed squared difference. To explore the different solutions that achieve the same squared error, we define $f_k^{min}$ and $f_k^{max}$ as the minimum and maximum values of $f^*_{\mathrm{k}}$ among the solutions to (5)-(6). These values can be obtained by solving K convex quadratic problems:

$$f_k^{min} = \min_{\mathrm{f}} f_k \tag{2.7}$$

subject to

$$\text{(2.6) and } ||Y - Qf||^2_2 = ||Y - Qf^*||^2_2 \tag{2.8}$$

and similarly for $f_k^{max}$. Solutions to these optimization problems are calculated using the CVX package [Grant and Boyd, 2013]. $f_k^{max}$ and $f_k^{min}$ provide bounds for the fractional values each cell type can achieve, in fact, for our transcriptional data, these bounds are widely separated, indicating a high degree of non-uniqueness of the solution. Moreover, all $f_k^{min}$ values computed are zero, indicating that no cell type is essential to fit this data (Figure 2.3). The NNCLS approach thus fails to provide a principled means of quantifying the uncertainty associated with each inferred cell type.

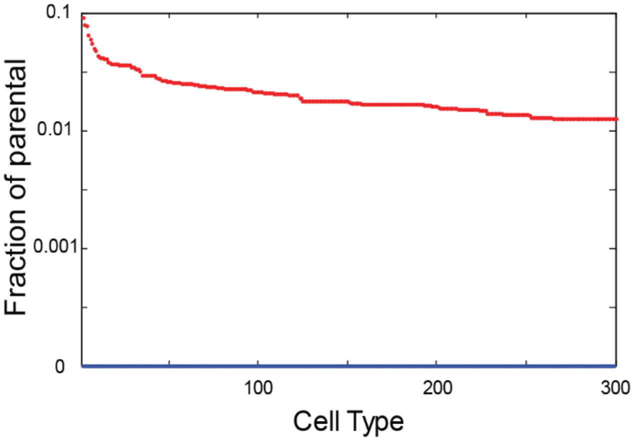


**Figure 23. Bounds on the fractional values achieved by the NNCLS solution computed using transcriptional information.**
Due to the non-uniqueness of the NNCLS solutions, each component can in general take many values while the overall solution maintains the same squared error. The minimum and maximum values of each fractional values ($f^{min}_k$ and $f^{max}_k$, respectively) are indicated by blue and red dots. Cell types are sorted in order of $f^{max}_k$ and we are only displaying the first 300 more important.

*Sparse Bayesian Formulation*

We therefore resorted to a Bayesian approach in which the unknown cell-type fractions are modeled as random variables, allowing their uncertainty to be characterized by a probability distribution. The use of a prior distribution enables previous knowledge and expectations to be incorporated into the model, and a likelihood function reflects the probability that the observed data were generated by the model. Combining the prior with the likelihood generates a posterior probability distribution that provides a level of confidence about the inferred results.

As a biologically plausible prior distribution over cell-type fractions, we chose to employ a constrained “spike-and-slab” distribution [Mitchell and Beauchamp, 1988]; [George and McCulloch, 1993]; [Ishwaran and Rao, 2005]. This prior incorporates the biologically reasonable assumption that only a small fraction of the 1,978 potential cell types will actually exist within the parental V1 population. Under this prior, many of the cell-type fractions are forced to be zero, with the consequence that only a small subset of the potential expression patterns is required to explain the measurements. The constrained spike-and-slab prior also enforces the obligate non-negativity and sum-to-one constraints on cell-type fractions. Combining prior and data likelihood using Bayes' rule results in a posterior distribution from which estimates of confidence about the existence and identity of cell types can be extracted.

More concretely, in order to obtain a sparse solution for the $f_k$, we introduce binary variables $b_k = 0, 1$ and continuous variables $z_k$ with prior distributions:

$$p\ (b_k \mid a) = \mathrm{a}^{bk}\ (1-\mathrm{a})^{1-bk}$$
$$p\ (z_k \mid \tau^2) = N(0, \tau^2)$$

(2.9)

such that $f_k = b_k\ \ z_k$ .The binary variable $b_k$ is responsible for the sparsity: with probability 1−a we have $b_k = 0$, which leads to $f_k = 0$. The variables $z_k$ are introduced for mathematical convenience and keep track of the values that $f_k$ can have whenever $b_k = 1$. Their prior distribution $N(0, \tau^2)$, reflects our beliefs about the values that non-zero $f_k$'s take. Hyperparameters $a$ and $\tau^2$ are treated as random variables and their prior distribution is considered flat, allowing their values to be determined by the data while exploiting the advantages of a Bayesian formulation. Since our noise model is Gaussian, the likelihood of the observed data y is quadratic on $f$:

$$p\ (Y \mid Q, f) \propto e^{-1\ 2f'Q'\,Q\ f\ +y'\,Q\,f}$$

(2.10)

Using Bayes theorem we can compute the log posterior distribution over $b, z, a, \tau^2$:

$$\log p(b, z, a, \tau^2 \mid Y, Q) = \log p\ (Y \mid Q, f) + \log p\ (b \mid a) + \log p\ (z \mid \tau^2) + \text{const.}$$

$$= -\tfrac{1}{2}\, f'\, M f + Y'\, Q\, f - \frac{z\, z}{2\, \tau 2} + |b = 1| \log(a) + |b = 0| \log(1-a) + \text{const.}$$

(2.11)

However, the posterior distribution cannot be computed directly, necessitating the use of a computational inference method based on Monte Carlo sampling [Gelman et al, 2013]. The Monte Carlo approach draws random samples from the posterior distribution. Given a large number of these samples, we can compute properties of the desired posterior distribution numerically, analogous to rolling a die repeatedly to determine the probability of each face appearing, rather than deriving that the probability is 1/6. To perform this computation we adapted a Hamiltonian Monte Carlo (HMC) algorithm that is specialized for constrained spike-and-slab posteriors and permits efficient sampling from our posterior distributions [Pakman and Paninski 2013, 2014].

*Sampling Algorithm*

Sampling from the posterior distribution gives us a computational method for inferring quantities of interest such as the posterior inclusion probability (confidence on the necessity of a cell type to explain the data), fractional values (with its mean and variance), etc. Since this is a complex distribution with many variables, we adopt a Gibbs sampling procedure [Gelman et al., 2013], which consists in successively sampling from the conditional distributions:

$$\mathrm{p}\ (a \mid b) \propto a\ |b = 1|\ (1-a)\ |b = 0| = \mathrm{Beta}(|b = 1| + 1,\ |b = 0| + 1)$$

$$\mathrm{p}\ (\tau^2 \mid z) \propto e^{-\frac{z\,z}{2\,\tau 2}} = \mathrm{InverseGamma}(1,\ z \cdot z/2)$$

$$\mathrm{p}\ (b, z \mid Y,Q,a,\tau^2) = -\ \tfrac{1}{2} f'\, M\, f + Y'\, Q\, f - \frac{z\,z}{2\,\tau^2} \tag{2.12}$$

The first two distributions can be sampled from using standard tools. The most difficult step is sampling from the mixed binary-Gaussian distribution, where we should impose the constraints:

$$z_k \geq 0 \quad , \quad \sum_k b_k z_k \leq 1 \quad , \quad \sum_k q_{i,k} b_k z_k > 0 \quad , \quad \text{as explained before.} \tag{2.13}$$

To sample from this distribution, we use the technique developed in [Pakman and Paninski, 2014, 2013], based on the Hamiltonian Monte Carlo (HMC). This is a Markov Chain Monte Carlo (MCMC) algorithm that uses ideas from particle dynamics to sample from complex probability distributions (see [Neal, 2010] for a review). The HMC algorithm proposes to write a Hamiltonian function in terms of the probability distribution we want to sample.

$$H(q, p) = U(q) + K(p) \tag{2.12}$$

Where U(q) is the potential energy, and will be defined as minus the log probability density of the probability distribution that we wish to sample, plus any constant that is convenient. Additionally, we must introduce auxiliary momentum variables, p. K(p) is called the kinetic energy, and is usually defined as

$$K(p) = ½\, p^t M^{-1} p \tag{2.13}$$

Writing the kinetic energy in this form corresponds to minus the log probability density (plus a constant) of the zero-mean Gaussian distribution with covariance matrix M. The HMC method alternates updating the momentum variables with Metropolis updates in which a new state is proposed by computing a trajectory according to Hamiltonian dynamics [Neal, 2010].

For constrained binary-Gaussian distributions, HMC is particularly efficient and explores the sampling space much faster than competing alternative approaches, such as Gibbs or random-walk Metropolis-Hastings. At the root of this efficiency lies a map of the binary variables b into continuous variables which leads to the dynamics of a particle in a piecewise-constant, constrained quadratic potential, which has exact analytical solutions (see [Pakman and Paninski, 2014, 2013] for details).

Briefly, to sample from the mixed binary-Gaussian distribution of 2.12, let’s explicitly write the prior using the spike-and-slab formalism,

$$
\begin{aligned}
p(fz_k\,|b_k\,,\tau^2) = \frac{1}{\sqrt[2]{2\,\pi\tau^2}}e^{-\frac{f_k{}^2}{2\tau^2}} & \qquad \text{for } b_k = +1,\\
\delta(f_k) & \qquad \text{for } b_k = +1,
\end{aligned}
\tag{2.13}
$$

We are interested in sampling from the posterior, given by:

$$
\begin{aligned}
p(f,b|Y,Q,a,\tau^2) &\propto p(Y,Q|f)p(f,b|a,\tau^2)\\
&\propto \frac{e^{-\frac{1}{2}f^+(M^+ + \tau^{-2})f^+ + \frac{YQf^+}{\sigma^2}}}{2\,\pi\tau^{2^{\frac{|b^+|}{2}}}}\delta(f^-)a^{|b^+|}(1-a)^{|b^-|}
\end{aligned}
\tag{2.14}
$$

Where subscripts + and − denote the subspace of active/inactive groups ($b_k$ = 1 or -1 respectively). To sample from this distribution more efficiently, we take two steps. First, we replace the delta function by a similar slab factor:

$$
\delta(f_k) \rightarrow \frac{1}{\sqrt[2]{2\,\pi\tau^2}}e^{-\frac{f_k{}^2}{2\tau^2}} \qquad \text{for } b_k = +1
\tag{2.15}
$$

Secondly, we augment the distribution with $y$ variables, such as $b = sign(y)$, and sum over $b$. This gives a distribution:

$$
p(f,b|Y,a,\tau^2) \;\propto \frac{e^{-\frac{1}{2}f^{+\prime}(M^+ + \tau^{-2})f^+ + \frac{YQf^+}{\sigma^2}}}{2\,\pi\tau^{2^{\frac{|b^+|}{2}}}}e^{-\frac{f^{-\prime}f^-}{2\tau^2}}e^{\frac{y\prime\, y}{2}}a^{|b^+|}(1-a)^{|b^-|}
$$

Where the values of $b$ in the rhs are obtained by taking sign(y). This is a piecewise Gaussian, different in each orthant of $y$, and possibly truncated in the $f$ space. Sampling from 2.15 gives us samples from the original distribution 2.11 using a simple rule: each pair $(z_k , y_k)$ becomes $(f_k , b_k = +1)$ if $y_k \geq 0$ and $(f_k = 0, b_k = -1)$ if $y_k < 0$. This reverse the transformations: the identification $b_k = \text{sign}(y_k)$ takes us from $p(f, y \mid Y, a, \tau^2)$ to $p(f, b \mid Y, a, \tau^2)$, and setting $f_k = 0$ when $b_k = -1$ reverse the slab transformation.

Using this posterior distribution as the potential energy and, introducing momentum variables, $g$, we can take advantage of the HMC method by writing a Hamiltonian as:

$$H = H_+ + H_- + H_y + H_a \tag{2.14}$$

$$\begin{aligned} H = \frac{1}{2} f'_+ \left(M_+ + \frac{1}{\tau^2}\right) f - y'Xf_+ + \frac{1}{2} g'_+ \left(M_+ + \frac{1}{\tau^2}\right) g_+ \\ + \frac{1}{2\tau^2} f'_- f_- + \frac{1}{2\tau^2} g'_- g_- \\ + y'y + q'q \\ + |b = 1| \log \frac{a}{1-a} + Const. \end{aligned} \tag{2.15}$$

Where again, subscripts + and − denote the subspace of active/inactive groups, g+, g−, q are momentum variables and $y$ is the variable expansion introduced to account for the dynamic of the $b_k$ variables. The algorithm alternates between sampling the momentum variables and then, using these sampled values and the last value of $f$ as initial conditions, sampling new values for $f$. The new value of $f$ is obtained by supposing that a particle is moving under the equations of motion defined by the Hamiltonian, which follows the following dynamics:

$$H = \frac{1}{2} f'Mf - rf + \frac{1}{2} G'MG$$
$$f_i(t) = \mu_i + a_i \sin(\text{t}) + b_i \cos(t)$$

$$\mu_i = \sum_j M_{ij} r_j$$
$$a_i = \dot{f}_i(0) = \sum_j {M_{ij}}^{-1}\, g_j\,(0)$$
$$b_i = f_i(0) - \mu_i$$

(2.16)

The trajectory of this auxiliary particle will continue until its position violates any constraint. These constrains represent boundaries imposed on the movement of the particle. In our problem, constraints are imposed by equation 2.13, enforcing two types of boundaries. The first boundary is violated when the particle reaches a region not allowed in state space, as in the event of violating $\sum_{\mathrm{k}} b_k z_k \leq 1$ and $\sum_{\mathrm{k}} q_{i,k}\, b_k\, z_k \;>\; 0$. In this case, the particle must bounce back at the boundary and reverse its momentum. This situation is equivalent to a particle bouncing against a wall.

Given a constraint satisfying a linear relationship, the general expression for the time at which a particle reaches boundary j (given by variables $K_j$ and $h_j$) is given by the solution of the following equations:

$$K_j\, f_j + h_j \geq 0$$
$$\sum_l K_l^j (\mu_l + a_j \sin(t) + b_j \cos(t)) h_j = 0$$
$$u_j \cos\left(t + \varphi_j\ \right) + \sum_l K_l^j \mu_l + h_j = 0$$
$$t = \operatorname{acos}\left(\frac{-K_j\,\mu - h_j}{u_j}\right) - \ \varphi_j$$

(2.17)

Where $u_j = \sqrt{\left(\sum_l K_l^j\ a_l\right)^2 + \left(\sum_l K_l^j\ b_l\right)^2}$ , $\varphi_j = \operatorname{atan}\left(-\frac{\sum_l K_l\, a_l}{\sum_l K_l\, b_l}\right)$. The coordinate violating that constrain should bounce and change direction. To calculate the change in velocity after the bounce, we will start by uncoupling the coordinates:

$$H = \frac{1}{2}(f-\mu)'M(f-\mu) + \frac{1}{2}\dot{f}'M\dot{f}$$
$$q = X(f-\mu)$$
$$H = \frac{1}{2}q'q + \frac{1}{2}\dot{q}'\dot{q}$$

(2.18)

Rewriting the constraint as:

$$K_j\, X^{-1}\left(q_j + \mu\right) + h_j \geq 0$$
$$K_j\, X^{-1}\dot{q}_j \geq 0$$
$$\widetilde{K}_j\, \dot{q}_j \geq 0$$

(2.19)

To obtain the reflected velocity, let's note that the vector *K<sub>j</sub>* is perpendicular to the reflecting plane *j* and, decomposing the velocity into a perpendicular and a parallel component to this place:

$$\dot{q}_{new} = \dot{q}_{old} - 2\alpha\widehat{K_J}$$

(2.20)

Where $\widehat{K_J} = \frac{K_j}{\|K_j\|}$, and $\alpha = \dot{q}_{old}\widehat{K_J}$.

$$\dot{q}_{new} = \dot{q}_{old} - 2(\dot{q}_{old}K_j)\frac{K_j}{\|K_j\|^2}$$

(2.21)

The second type of boundary is met when a variable $y_k$ crosses an octant ($y_k = 0$), changing the estate of a fractional value $f_k$ from being active to inactive (or vice versa). During this transition, the conservation of energy requires that:

$$\frac{\dot{q}_j^2(t^+)}{2} = \Delta_j + \frac{\dot{q}_j^2(t^-)}{2}$$

(2.22)

and the energy jump $\Delta_j$ depends on *f* and is given by:

$$\delta H = \delta H_+ + \delta H_- + \delta H_y + \delta H_a = 0$$

$$\delta q = \delta H_+ + \delta H_- + \delta H_a$$

$$\Delta_j = -H_{f,q}\,(b_{-k}, b_k = +1) + H_{f,q}(b_{-k}, b_k = -1) + \log\left(\frac{a}{1-a}\right) \tag{2.23}$$

If 2.23 results in a positive balance for the kinetic energy, the particle crosses the $y_j = 0$ boundary, and if not, it bounces back with $q_j(t_{+j}) = -q_j\,(t_{-j})$.

*Wang Landau algorithm to improve sampling*

Hamiltonian Monte Carlo algorithms have difficulties moving between regions in the posterior distribution of low probabilities. Imposing constraints creates isolated modes in our posterior distribution, which in practice slow the mixing of the HMC Markov chain: it is difficult for the particle to move from one region to another. This is caused when a non-active component ($b_k = 0$) can only become active ($b_k = 1$) when it does not violate any of the inequalities 2.13. But for typical values of $z_k$ (dictated by $\tau^2$), this is not a usual situation. A useful technique to encourage an MCMC algorithm to better explore isolated regions is simulated tempering [Marinari and Parisi, 1992]. In this method, we introduce an additional discrete random variable $t$ that takes values in $t = 1,...,$T and parametrizes 'inverse temperatures' $\beta_t$, and augment the distribution of interest as:

$$p(f, b, a, \tau^2, t|Y, Q) = p(f, b, a, \tau^2|Y, Q, t)p(t) \tag{2.24}$$

Where

$$p(z, b, a, \tau^2|Y, Q, t) = \frac{\tilde{p}(z,b,a,\tau^2|Y,Q)^{\beta t}}{\theta_t} \text{ and } p(t) = \frac{1}{t}. \tag{2.25}$$

The distribution $\tilde{p}(z, b, a, \tau^2|Y, Q)$ is not normalized and the constants $\theta_t$ guarantee the proper normalization. Constraints 2.13 hold for all the values of t. The idea is to build a Markov chain that samples from (2.24) and keep only those samples with $t = 1$. For distributions with isolated maxima, one usually considers $\beta_1 = 1$ and $\beta_{t+1} < \beta_t$. These values make the distributions

(2.25) flatter for higher *t*, so when the Markov chain visits higher temperature states it can jump between maxima before returning to $t = 1$. But our case is different because, as mentioned, the exploration of different regions is hindered by the big values taken by $z_k$. The solution is therefore to lower the temperature, causing the distribution of inactive $z_k$'s to peak around zero, thus taking smaller values and allowing more components to become active. In order for the Markov chain to explore each temperature *t* with equal probability, we need the values of $\theta_t$ in (2.25). Although a direct computation is difficult, we can use the Wang-Landau algorithm [Wang and Landau, 2001]; [Atchad´e and Liu, 2010], which updates the values of $\theta_t$ based on the visited values of t to achieve a uniform rate of visits for all temperatures. See Algorithm 1 for details. In summary, a typical run of the Wang – Landau algorithm explore much more efficiently the posterior by inducing fast mixing (Figure 2.4).

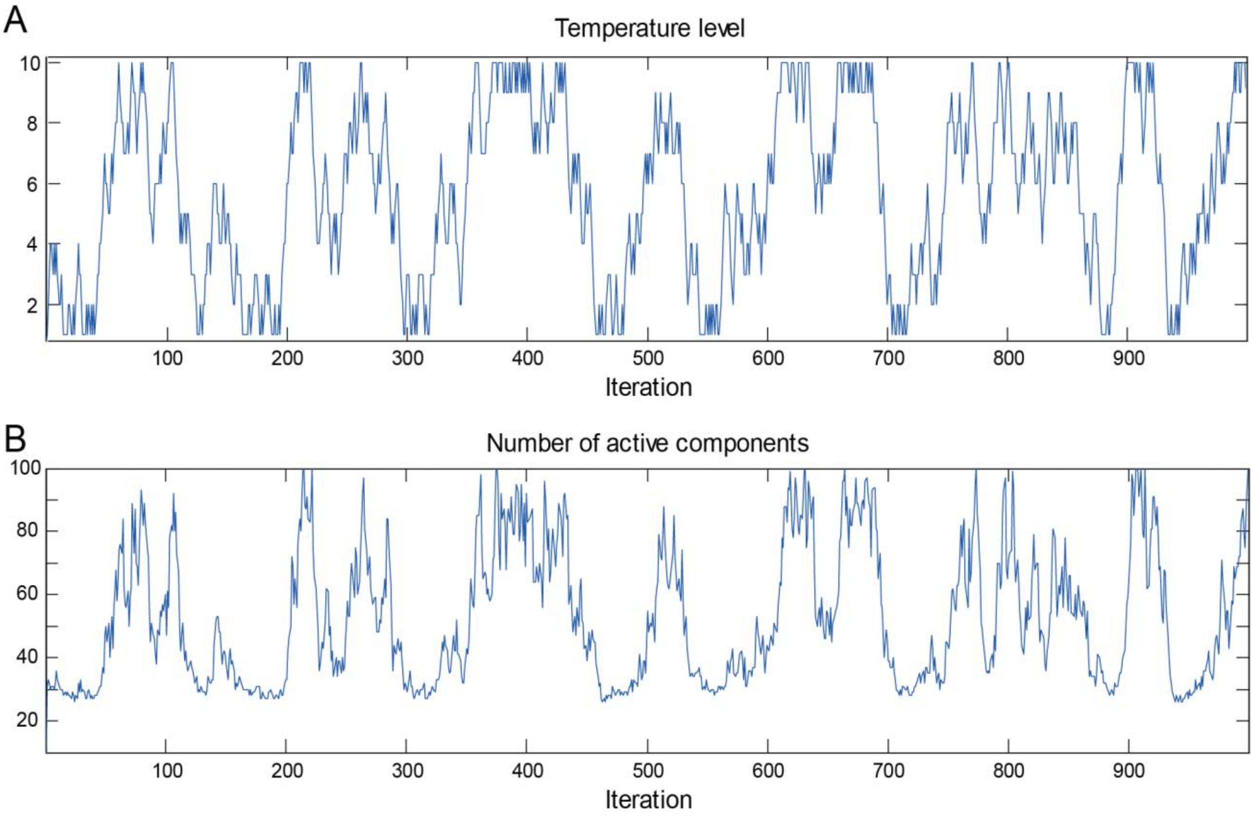
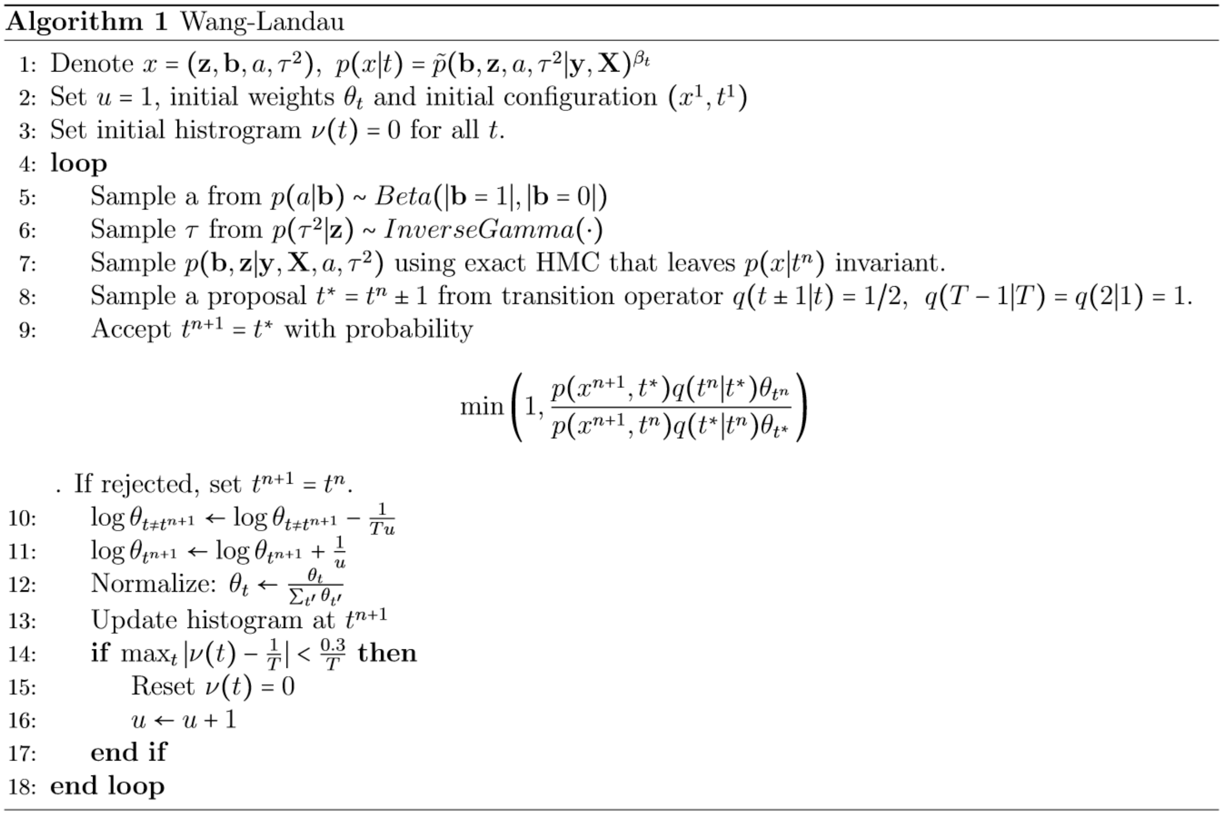


**Figure 2.4: Simulated tempering facilitates exploration of the diversity landscape.** Temperature levels and number of active components in 1000 iterations of the simulated tempering Markov chain. Inverse temperatures is equally spaced in ten intervals between β1=1 and β10=1.2. Note the strong correlation between low temperatures and high number of active components.

(a) Temperature level at different iterations.

(b) Number of active components for each iteration.



```
Algorithm 1 Wang-Landau
Denote $x = (\mathbf{z}, \mathbf{b}, a, \tau^2)$, $p(x|t) = \tilde{p}(\mathbf{b}, \mathbf{z}, a, \tau^2|\mathbf{y}, \mathbf{X})^{\beta_t}$
Set $u = 1$, initial weights $\theta_t$ and initial configuration $(x^1, t^1)$
Set initial histrogram $\nu(t) = 0$ for all $t$.
loop
    Sample a from $p(a|\mathbf{b}) \sim Beta(|\mathbf{b} = 1|, |\mathbf{b} = 0|)$
    Sample $\tau$ from $p(\tau^2|\mathbf{z}) \sim InverseGamma(\cdot)$
    Sample $p(\mathbf{b}, \mathbf{z}|\mathbf{y}, \mathbf{X}, a, \tau^2)$ using exact HMC that leaves $p(x|t^n)$ invariant.
    Sample a proposal $t^* = t^n \pm 1$ from transition operator $q(t \pm 1|t) = 1/2$, $q(T - 1|T) = q(2|1) = 1$.
    Accept $t^{n+1} = t^*$ with probability
        $$\min\left(1, \frac{p(x^{n+1}, t^*)q(t^n|t^*)\theta_{t^n}}{p(x^{n+1}, t^n)q(t^*|t^n)\theta_{t^*}}\right)$$
    . If rejected, set $t^{n+1} = t^n$.
    $\log\theta_{t\neq t^{n+1}} \leftarrow \log\theta_{t\neq t^{n+1}} - \frac{1}{Tu}$
    $\log\theta_{t^{n+1}} \leftarrow \log\theta_{t^{n+1}} + \frac{1}{u}$
    Normalize: $\theta_t \leftarrow \frac{\theta_t}{\sum_{t'}\theta_{t'}}$
    Update histogram at $t^{n+1}$
    if $\max_t |\nu(t) - \frac{1}{T}| < \frac{0.3}{T}$ then
        Reset $\nu(t) = 0$
        $u \leftarrow u + 1$
    end if
end loop
```

**Algorithm 2.1. Sampling algorithm.**

*Sampling Procedure*

Each iteration of the sampling algorithm generates a set of cell-type fractions that satisfy the constraints and provide a good fit to the data. Generated in this way, the number of selected cell types and their expression patterns vary across iterations. Multiple samples, collected across a large number of iterations, allow us to infer the properties of the posterior probability distribution. For example, the proportion of Monte Carlo samples for which a particular expression pattern is selected (i.e., for which the corresponding cell-type fraction is non-zero) determines that type's posterior inclusion probability, which serves as a confidence measure of its necessity to explain the data. We can also compute the distribution of the number of cell types selected in each iteration: this allows us to estimate the total number of distinct cell types required to explain the observed data. Repeated sampling also enables us to compute cross-correlations between cell-type fractions; these are used to construct a list of candidate expression profiles along with the probabilities of their corresponding to actual cell types. As with an expression pattern, a candidate expression profile is a 19-component vector, with each component representing a different transcription factor, but now these components are allowed to be real numbers between 0 and 1, instead of binary. Component a of any candidate expression profile represents the probability that transcription factor *a* is expressed as part of that profile.

### *Computational Validation Experiments*

We validated the ability of the Bayesian approach to accurately infer cellular diversity by performing computational cross-validation experiments, as well as experiments on simulated datasets, for which the underlying cell types and corresponding cell-type fractions are known. Here, we reproduced a leave-one-out cross-validation test implemented by running the algorithm on datasets obtained by successively leaving out the measurements corresponding to one factor or pair of factors. The predicted values for the fraction of left-out factors can be compared to measured values. Almost all predicted values lie within 2 standard deviations from the measured data, showing that we can estimate these values correctly. Verification of the ability of the methods to recover meaningful and accurate estimates of cellular diversity permitted us to apply these algorithms to V1 transcriptional datasets.

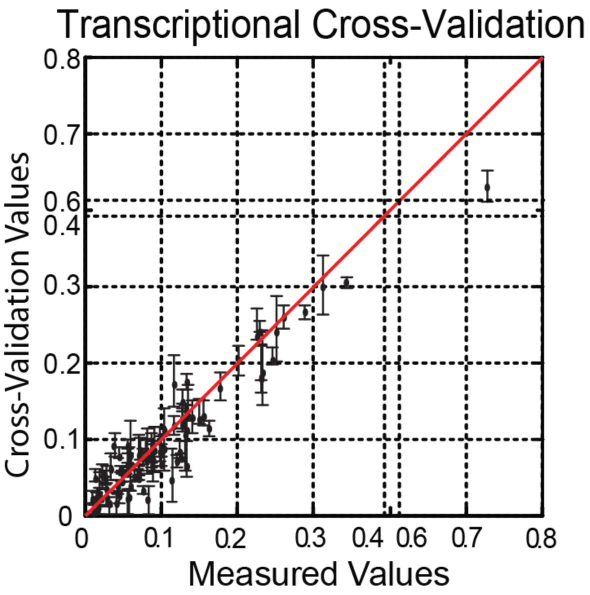


**Figure 2.5: Cross-Validation experiments.** Leave-one-out cross validation experiment showing good agreement between measured and cross-validated values.

## 2.2.3 V1 diversity extracted from transcriptional data alone

We applied this Bayesian analysis to transcriptional information as outlined above (Figure 2.2a and 2.2b). As discussed above, each iteration of the HMC sampling algorithm generates a possible set of cell types, but their number and identity vary across HMC iterations. Over the course of the full HMC run, the number of types selected (those with non-zero cell-type fractions) ranged from 25 to 33 with a mean ± standard deviation of 29 ± 2 (Figure 2.6a). The transcriptional identity of the selected cell types varied widely across different HMC iterations. For example, if two iterations each resulted in 30 selected cell types, then typically some of the cell types chosen in the first iteration would not appear in the second, and vice versa.

Computing the posterior inclusion probability of each expression pattern across many samples led to a rank-ordered list of candidate expression patterns. The 40 candidate patterns with the highest inclusion probabilities (i.e. those that appeared most frequently in the HMC samples) and their inferred cell-type fractions are shown in Figure 2.6b. The expression pattern with the highest inclusion probability corresponds to the Renshaw interneuron, a V1 neuronal type that mediates recurrent inhibition of motor neurons [Renshaw, 1946] and co-expresses the transcription factors Onecut1, Onecut2 and MafB [Stam et al, 2012]. This analysis also infers the existence of MafA+ and MafA- subsets of Renshaw interneurons (patterns 1 and 30 in Figure 2.6b), a molecular diversity that may correspond to the known morphological heterogeneity exhibited by Renshaw interneurons [Fyffe, 1990].

We next addressed the sensitivity of these results to the number of transcription factors used in the analysis, wondering if the incorporation of additional TFs would dramatically expand cell-type diversity. To assess this possibility we evaluated the extent of diversity that emerges from the use of subsets of only 11 to 18 of the 19 measured transcription factors. The average number of selected cell types – a value of 29 for the analysis using the full 19 factors - tends to decrease gradually when smaller numbers of transcription factors are analyzed. Thus when 16 to 19 transcription factors were incorporated, the number of selected cell types remains relatively constant and close to 29 (Figure 2.6c). In contrast, the number of potential cell types (1,978 for the case of 19 factors) depends much more strongly (over almost an order of magnitude) on the chosen transcription factor subset (Figure 2.6d). These findings suggest that Bayesian calculations of cellular diversity based solely on transcription factor data may already be close to saturating with the use of 19 transcription factors.

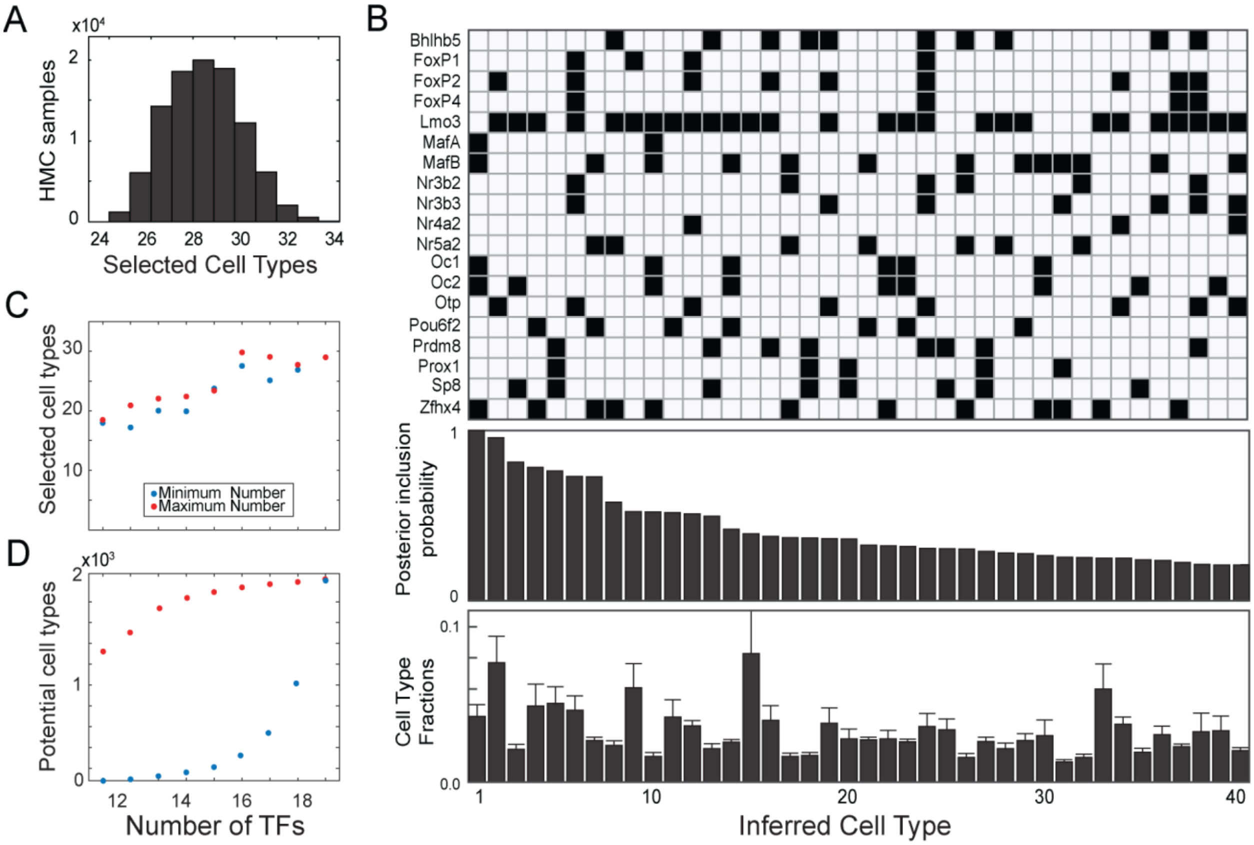


**Figure 2.6. Cell Type Discovery using Transcription Factor Expression Information**

(a) Number of cell types selected per HMC iteration (for which the fraction $f_k$ was nonzero).

(b) Transcriptional profiles of top 40 inferred cell types. Cell types (top) are arranged by descending posterior inclusion probability (middle). Black indicates TF expression, white indicates absence of expression. Bottom: fraction of each cell type in the parental V1 population (mean ± SD of all nonzero sampled values).

(c) Number of selected cell types remains close to 29 when varying the set of observed TFs. Red and blue curves denote the maximum and minimum number for different TF sets.

(d) Number of potential cell types. Red and blue curves denote maximum and minimum numbers after reduction by measured TF pairs that exhibit no co-expression.

## 2.2.4 Clustering Cell Types into Groups

What is the diversity of potential transcriptional expression patterns within the parental V1 population? We detected 131 different expression patterns with posterior inclusion probabilities greater than 0.05 (i.e. appearing in more than 5% of the HMC samples), of which only the 40 expression patterns appearing most frequently in the HMC samples are indicated in Figure 2.6b. The existence of 131 candidate expression patterns for only ~30 cell types (the average number selected in each sample; Figure 2.6a), and the fact that few of the top expression patterns in Figure 2.6b have posterior inclusion probabilities near one, is another indicator of the incompleteness of the expression-only data for specifying cell-type identity, and thus the nonuniqueness of the resulting solution (recall the NNCLS solution).

We constructed *candidate expression profiles* by clustering the 131 most likely expression patterns into ‘groups’. A group is defined as a set of expression patterns that satisfies two conditions: (i) the members of a group express similar sets of TFs (Figure 2.7a), and (ii) in all or almost all of the HMC samples, only a single member of a group is selected (i.e. has a non-zero cell-type fraction), although different members may be selected in different samples (Figure 2.7b). The second condition causes the members of a group to be negatively correlated with each other (Figure 2.7a, table on the right). These conditions imply that a group is likely to represent a single cell type with an uncertain expression pattern, rather than multiple cell types.

We developed a recursive algorithm for constructing such groups. All candidate expression profiles with inclusion probabilities greater than 5% were assigned to groups, with most groups having only a single member selected across all of the HMC samples, and no group having more than one member selected in >3% of the samples (Figure 2.7d). To examine the robustness of the inferred groups, we systematically varied the threshold for selecting the list of candidate expression patterns from which groups were constructed. As this threshold is lowered, the number of groups first increases linearly because each high-ranked expression pattern spawns its own group (Figure 2.7c). However, this growth slows as lower-ranked patterns join existing groups. The result is that the number of candidate groups depends only weakly on the inclusion threshold, once the threshold is sufficiently low and sufficient expression patterns are included. With an inclusion threshold of 5%, the clustering algorithm identifies 35 groups (Figure 2.7e-f).

Each group gives rise to a single candidate expression profile (Figure 2.8a-b), and for each profile we assign an expression probability to each transcription factor, weighting the binary expression patterns of each member of the group by the frequency with which it appear in the HMC samples (Figure 2.8b, top). In addition, we compute a posterior inclusion probability (Figure 2.8b, bottom) for each candidate expression profile, which gives the probability that the expression profile should be designated one of the inferred cell types. The posterior inclusion probability of several of the candidate expression patterns is close to one, much higher than the inclusion probabilities of the corresponding candidate expression patterns from which they are constructed (Figure 2.6b). Nevertheless, there is still considerable uncertainty in the identity of the cell types predicted by the transcriptional data alone (Figure 2.8b).

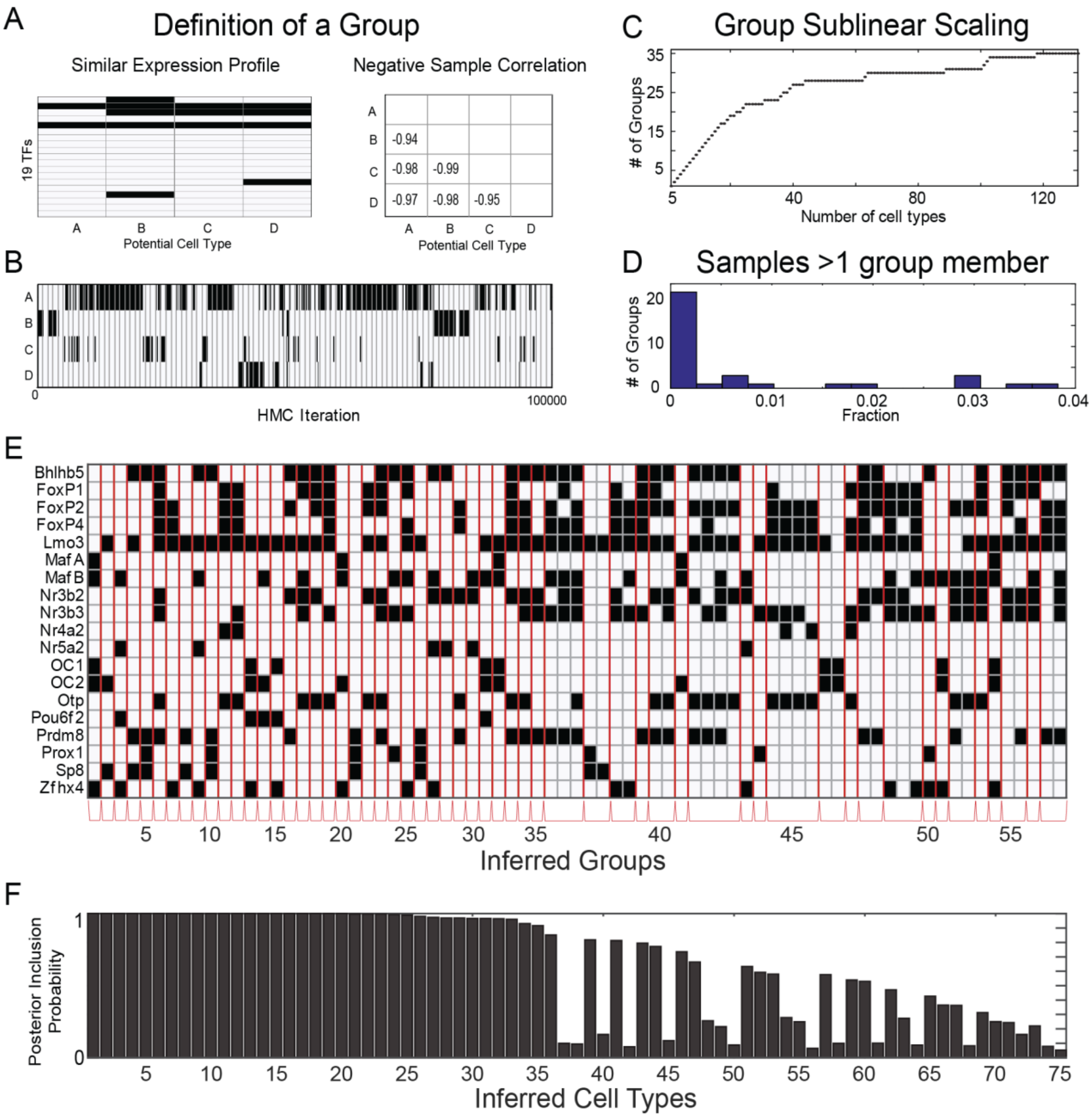

A
Definition of a Group
Similar Expression Profile
19 TFs
A
B
C
D
Potential Cell Type
Negative Sample Correlation
A
B
C
D
-0.94
-0.98
-0.99
-0.97
-0.98
-0.95
A
B
C
D
Potential Cell Type
B
A
B
C
D
0
100000
HMC Iteration
C
Group Sublinear Scaling
# of Groups
35
25
15
5
5
40
80
120
Number of cell types
D
Samples >1 group member
# of Groups
20
10
0
0
0.01
0.02
0.03
0.04
Fraction
E
Bhlhb5
FoxP1
FoxP2
FoxP4
Lmo3
Maf A
Maf B
Nr3b2
Nr3b3
Nr4a2
Nr5a2
OC1
OC2
Otp
Pou6f2
Prdm8
Prox1
Sp8
Zfhx4
5
10
15
20
25
30
35
40
45
50
55
Inferred Groups
F
Posterior Inclusion Probability
1
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
Inferred Cell Types

**Figure 2.7 (preceding page). Definition of clustering algorithm that creates groups of correlated cell types.**

(a) Methodology used to classify cell types into V1 groups. Left, similar expression profiles: Transcription factors expressed by members constituting a particular transcriptional only group (group 18, Figure 2). Right, negative sample correlation: cross-correlation values calculated from the time series corresponding to the presence or absence of each group member in the set of selected cell types (values correspond to group 18).

(b) 100,000 iterations of the HMC sampling algorithm, demonstrating the presence (black) or absence (white) of members of group 18. In each iteration, generally only one of the four possible members is chosen.

(c) The number of groups scales sublinearly as a function of the number of underlying inferred cell types, indicating that increasing the number of candidate expression profiles has a relatively small effect on the number of inferred groups.

(d) Most of the members of transcriptionally defined groups are mutually exclusive. Histogram depicting the number of transcriptional groups, sorted by the fraction of samples in which more than one member of the group was selected. For 23 groups, only one group member was ever selected in a given iteration; for all remaining groups, more than one group member is selected in less than 3.5 percent of the samples. Similar results were obtained for spatially-defined groups (not shown).

(E) Transcription factors expressed by transcriptionally and spatially defined cell types with a posterior inclusion probability greater than 5 percent. Group members are arranged in between red lines.

(F) Posterior inclusion probability for cell types in (E). Cell types are ranked by group-level inclusion probability.

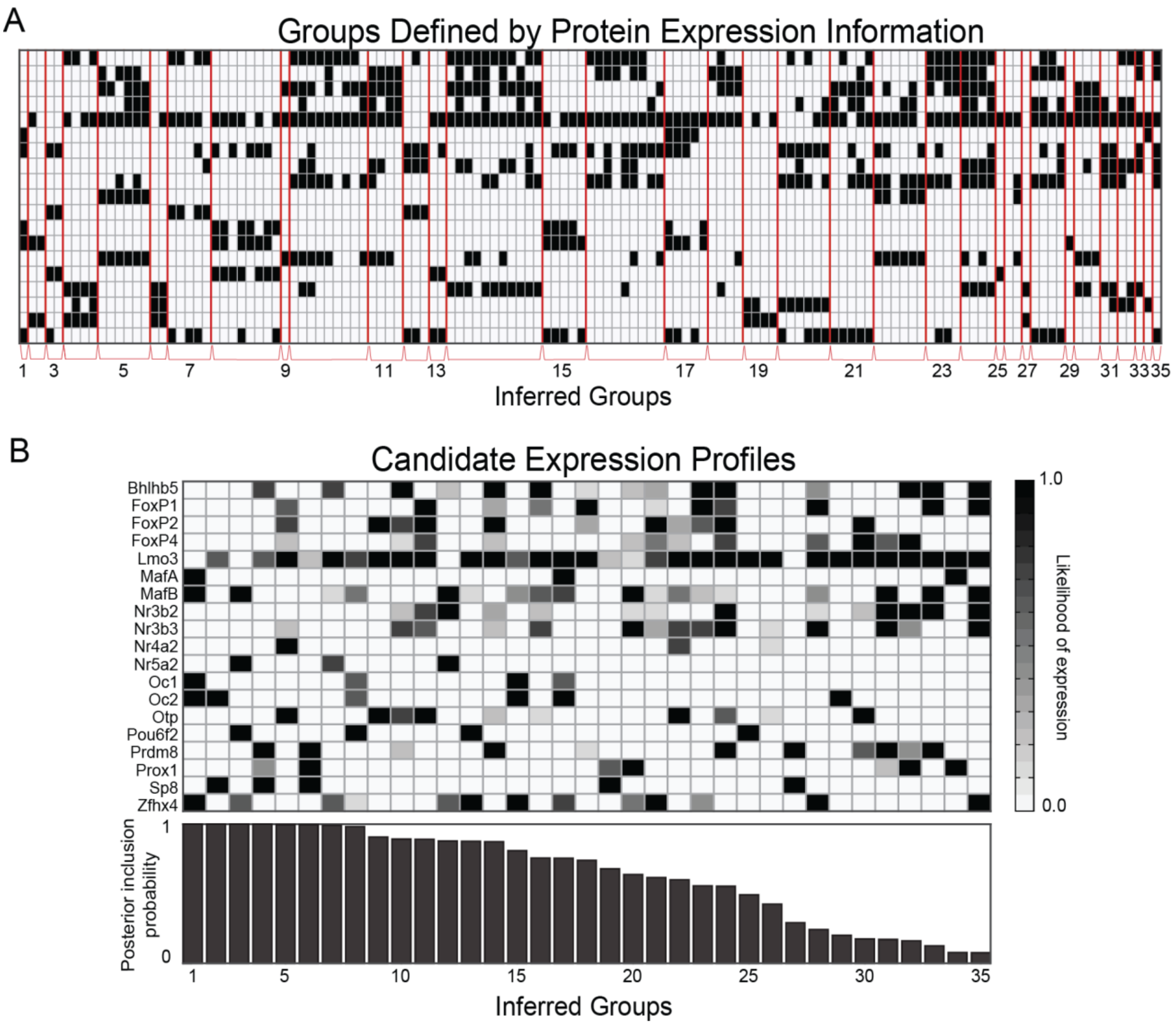


**Figure 2.8. Clustering Algorithm Arranging Cell Types into Correlated Groups**

(a) TFs expressed by cell types with a posterior inclusion probability greater than 5 percent. Inferred group members are arranged in between red lines.

(b) Representation of inferred groups. Top: candidate expression profiles derived from the 35 V1 groups. Gray scale indicates the likelihood of each TF expressed within the group. Bottom: posterior inclusion probability for each V1 group.

## 2.3 Discussion

Spinal interneurons shape motor activity and limb movement, but the organizational logic of their encoded microcircuits has remained obscure. By focusing on V1 interneurons we identified nineteen transcription factors that delineate extensive diversity within this inhibitory set. In addition, we presented an in-depth molecular characterization using the aforementioned transcription factors in which diversity is estimated through the use of a statistical procedure that infers underlying cell types by means of a sparse Bayesian algorithm.

Previous statistical models to extract cell type-specific information from gene expression data of heterogeneous populations formulated the problem by assuming that measured gene expression levels $e_i$ are a linear combination of underlying cell types $e_i = \sum_k E_{i,k} \; f_k + \varepsilon_i$ (see [Shen-Orr and Gaujoux, 2013] for a review). When the expression profiles of the cell types, $E_{i,k}$ are known, this equation can be solved for the population fractions $f_k$ [Lu et al., 2003]; [Gong et al., 2011]; [Grange et al., 2014]. Likewise, when the fractions $f_k$'s are known, one can solve for the gene expression level profiles. Finally, when only the measurements $e_i$ are known, factor analysis/matrix factorization methods can be used to find both $E_{i,k}$ and $f_k$ [Erkkila¨ et al., 2010]; [Gaujoux and Seoighe, 2012]; [Bazot et al., 2013]. However, as the case of the NNCLS approach, the solutions are generally non-unique and extracting biological information can be challenging without further assumptions. The last consequence is the product of a mathematical indeterminacy in which inferred fractional values are not uniquely defined, an affine transformation can be applied to them and, the inverse to their candidate expression profiles. These transformed fractional values would remain being a solution of the matrix factorization problem but their significance would be difficult to understand.

Our method resolves the previous challenges by considering population fractions and instead of inferring cell type profiles and fractional values simultaneously, a binary matrix $J_{k,I}$ is used, which enumerates every possible cell type consistent with the data. This expansion results beneficial permitting the usage of all cell types to explain the data, selecting only the necessary ones. Additionally, the inferred values $f_k$ have a clear biological meaning representing cell type fractions. Lastly, by using a Bayesian methodology in which sampling is performed to achieve inference, we gain the ability to assign confidence interval to each inferred quantity.

The distinctive feature of our approach, the enumeration all of the possible expression patterns, $2^{19}$ patterns in our case for the 19 genes considered, extends in the general case to $2^N$ for a study involving N genes. We note that this factor $2^N$ may be prohibitive for applications of the method to RNA-seq data, where a large number of genes are typically tracked. Although applications in which N is several thousand would appear impractical, it may be possible to identify a subset of genes expected to be particularly informative about cell type and restrict the analysis to this subset. Even with a reduced N, $2^N$ may be dauntingly large, but it is important to recall that in our analysis a preliminary screening reduced the number of expression patterns by a factor of ~265, from $2^{19}$ (524,288) to 1,978. Greater N values may yield even larger reductions.

Our analysis has identified extensive transcriptional diversity within V1 interneurons on the basis of the expression of 19 TFs. The first issue this raises is whether further diversity will follow inevitably with the inclusion of additional V1 TFs. We analyzed the impact of varying the number of TFs in our analysis and found only a weak dependence of the number of cell types on the number of TFs (Figure 2.6c-d). Thus, 19 TFs appear sufficient to uncover a substantial fraction of the total underlying transcriptional heterogeneity.

Is it possible to characterize V1 interneurons using marker information at the pairwise level? Candidate expression profiles and their prevalence within the parental population have been inferred from this data. Nonetheless, much uncertainty remains in the assignment of each candidate expression profile. To resolve this issue, inferred cell types have to be looked under the glass of the group assignment. By considering grouped candidate expression profiles, we gain confidence in the existence of each cell type by exchanging indeterminacy in the expression of each transcription factor within each profile. Otherwise, more information have to be provided to constrain inference. This is the purpose of the next chapter in which information about single transcription factor settling position grants us the ability to inferred more accurate candidate expression profiles.

# Chapter 3

# Spatial organization of V1 interneuron subpopulations

## 3.1 Introduction

Despite advances in elucidating the wiring of spinal motor systems, the organization of local circuit interneurons remains obscure. In much the same way that limb-innervating motor neurons acquire diverse pool identities, we showed in the previous chapter that cardinal interneuron classes defined by transcriptional character fragment into multiple types. Nonetheless, our inference algorithm produced estimates of cell diversity with high uncertainty.

To gain insight into the organization of inhibitory circuits for motor control, in this chapter we analyze diversity within the V1 interneuron population in light of information relating settling position to transcriptional character. The relevance of neuronal settling position in spinal connectivity has emerged from studies on the synaptic organization of sensory connections with motor neurons. Proprioceptive afferents target distinct dorsoventral domains of the ventral spinal cord in a manner independent of motor neuron character (Sürmeli et al., 2011), and thus the stereotypy of settling position is needed for the formation of selective sensory connections. Moreover, different physiological subtypes of interneurons appear to occupy stereotypic settling positions within the intermediate and ventral spinal cord [Alvarez and Fyffe, 2007]; [Hultborn et al., 1971]. However, whether interneuronal microcircuits use stereotypic patterns of neuronal position when establishing local micro circuitry remains unclear.

## 3.2 Results

### 3.2.1 Settling position of subsets of V1 interneurons

The notion that V1 transcriptional heterogeneity reflects functional distinctions in interneuron identity raises the possibility that V1 cell types are clustered in stereotypic settling patterns, akin to the spatial order of spinal motor neuron pools [Sürmeli et al., 2011] or the discrete domains occupied by certain classes of spinal interneurons [Thomas and Wilson, 1965]; [Hultborn et al, 1971]. Since individual V1 cell types are defined by as many as 9 TFs (Figure 2.6b), it is not practical to delineate them on the basis of their complete transcriptional profile. We therefore assessed the spatial distributions of V1 subpopulations defined by each of the 19 V1 TFs individually (Figure 3.1a) or in a few cases by the superimposition of two TFs (Figure 3.1b). These larger sets of V1 interneurons are each predicted to contain multiple V1 cell types (indicated from the predictions obtained in Figure 2.6b). Nevertheless, any spatial restriction of these larger V1 interneuron sets indicates clustering of individual V1 types.

We first examined a single case in which an inferred V1 cell type can be delineated by the co-expression of just two TFs, Nr5a2 and Pou6f2 (Figure 2.8). The spatial restriction of this $V1^{Nr5a2/Pou6f2}$ cell type revealed a highly stereotyped and localized settling position, with respect to the parental V1 population, which extends ~400 µm along the dorsoventral and mediolateral axes in p0 lumbar spinal cord (Figure 3.1b). Further analysis of the distributions of groups of V1 cell types revealed that each occupied a domain more restricted than that of the parental V1 distribution profile (Table 3.1, $p < 0.00001$ by one-tailed Monte Carlo test; single TF-gated fractional area (Fa) range: 0.217 to 0.855, dual TF-gated Fa range: 0.085 to 0.365, Figure 3.2a-b). By extension, it follows that the individual V1 cell types contained within these larger populations must also be clustered.

We also examined the degree to which subsets of V1 interneurons settle at distinct positions along the mediolateral or dorsoventral axes of the ventral spinal cord, focusing initially on four specific populations that highlight because of their mutual exclusiveness. The $V1^{MafA}$, $V1^{Pou6f2}$, and $V1^{Sp8}$ populations showed discrete distributions along the dorsoventral and mediolateral axes (Figure 3.2c-d and Table 3.1), whereas the $V1^{FoxP2}$ population occupies a broader spatial domain (Figure 3.1a, Fa = 0.855). Moreover, $V1^{Pou6f2}$ interneurons fractionated into medial Nr5a2+ and lateral Lmo3+ interneurons, with 93% mutual exclusion, revealing positional segregation within the members of a single population (Figure 3.2e). A similar segregation along the dorsoventral axis was seen within the $V1^{Prdm8}$, which can be fractionated into dorsal Sp8+ and ventral FoxP4+ subsets (Figure 3.2f). Importantly, analysis of $V1^{Sp8}$ and $V1^{Pou6f2}$ populations demonstrated that V1 settling positions are stable from e15.5 to p20, indicating that V1 neuronal clustering is not a transient reflection of developmental maturity (Figure 3.3a-b).

We also examined whether neurons in $V1^{Pou6f2}$ and $V1^{Sp8}$ exhibit rostrocaudal distinctions in settling position, prompted by the observation that Pitx2+ V0 interneurons exhibit rostrocaudal variation in transmitter phenotype and connectivity along the lumbar spinal cord [Zagoraiou et al., 2009]. The overall positional bias of $V1^{Pou6f2}$ and $V1^{Sp8}$ interneurons was maintained within the parental V1 domain (Figure 3.3c-e). Moreover, the mean position of V1 cell types and groups of cell types at L3-L5 levels of lumbar spinal cord was consistent between animals (Figure 3.2i).

Nevertheless, we detected minor differences in the settling position of V1$^{Pou6f2}$ and V1$^{Sp8}$ populations, with a ventromedial shift for V1$^{Pou6f2}$ and a ventrolateral shift for V1$^{Sp8}$ at progressively more caudal lumbar levels (Figures 3.2h and 3.3d-e). Thus, V1 interneuron settling position exhibits overall constancy, but with subtle differences along the lumbar rostrocaudal axis.

In total, this analysis of V1 identity and settling position identifies numerous spatially discrete V1 subpopulations, seven of which are illustrated (Figure 3.2i). These seven clusters represent ~44% of the parental V1 population. Collectively, these findings indicate that transcriptionally distinct V1 subpopulations exhibit a high degree of clustering and distinct spatial structures. They indicate further that the extent of V1 interneuron diversity goes beyond that recognized previously within this, or any other, interneuron population in the mammalian CNS [Sanes and Masland, 2015].

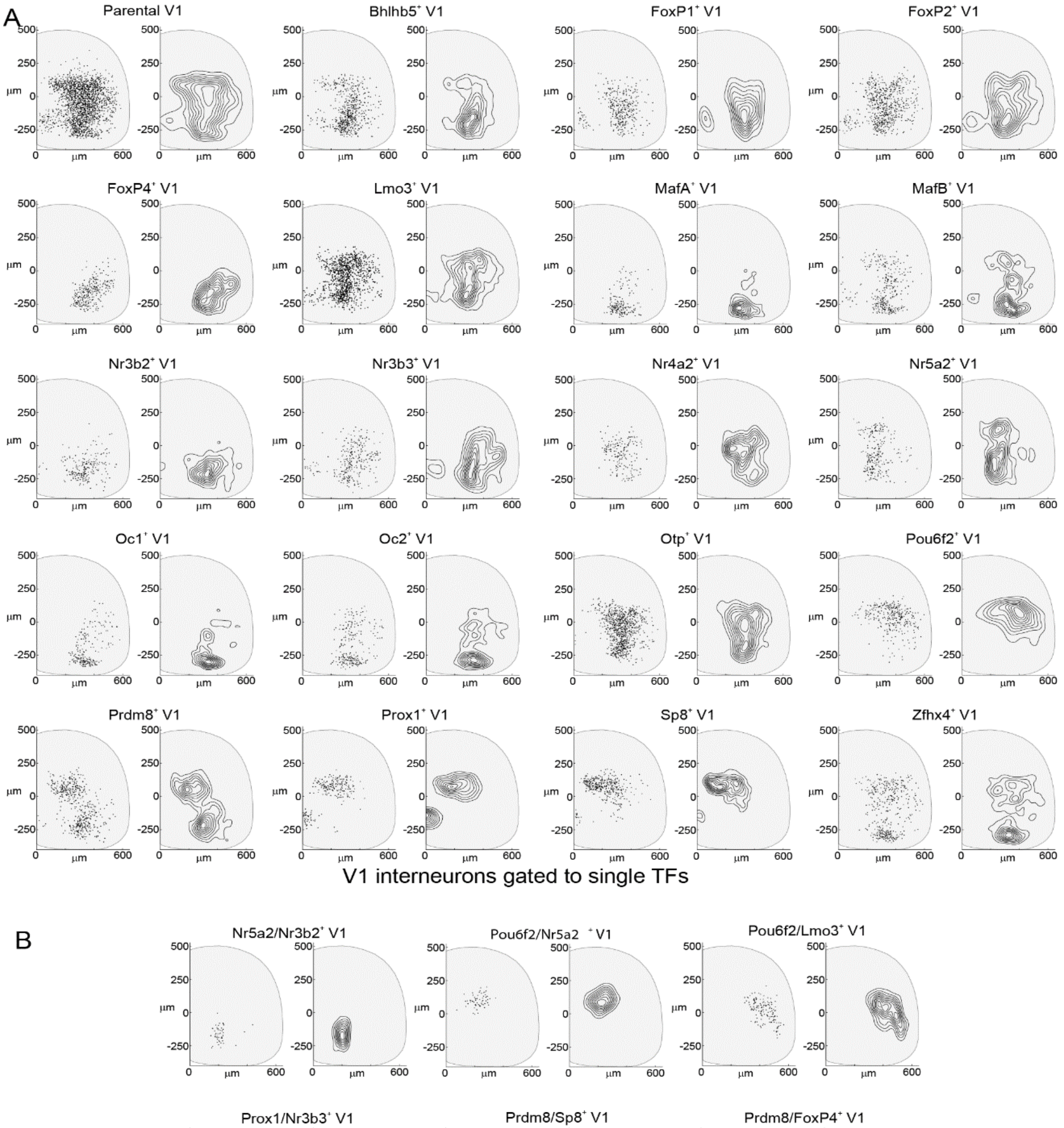


**Figure 3.1. Spatial Distributions of Transcriptionally Defined V1 Subpopulations.**
(a) Summary of the spatial distributions of parental V1 interneurons and V1 subpopulations gated to single transcription factors in p0 L3-L5 spinal segments. Each of these subpopulations contains multiple inferred V1 cell types, and is therefore termed a "composite group". Left panels indicate the position of individual interneurons, while right panels show density contours (10th-90th percentile kernel density estimates).

(b) Analogous spatial distributions for V1 subpopulations gated to two TFs. Pou6f2/Nr5a2 denotes a TF combination that uniquely defines a single inferred V1 cell type.

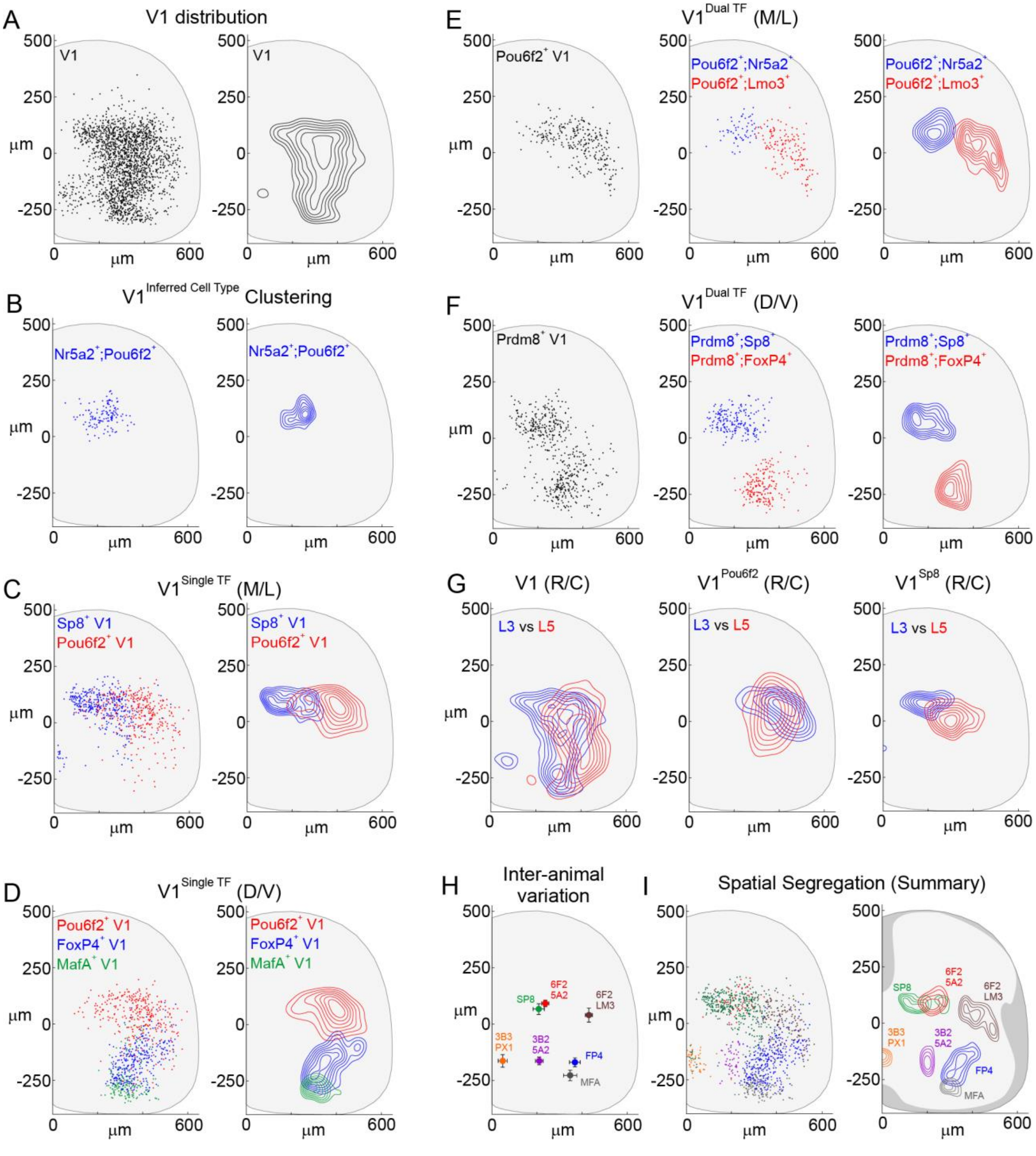

A
V1 distribution
V1
V1
500
250
0
-250
μm
0
μm
600
B
$V1^{\text{Inferred Cell Type}}$ Clustering
$Nr5a2^{+};Pou6f2^{+}$
$Nr5a2^{+};Pou6f2^{+}$
C
$V1^{\text{Single TF}}$ (M/L)
$Sp8^{+}$ V1
$Pou6f2^{+}$ V1
$Sp8^{+}$ V1
$Pou6f2^{+}$ V1
D
$V1^{\text{Single TF}}$ (D/V)
$Pou6f2^{+}$ V1
$FoxP4^{+}$ V1
$MafA^{+}$ V1
$Pou6f2^{+}$ V1
$FoxP4^{+}$ V1
$MafA^{+}$ V1
E
$V1^{\text{Dual TF}}$ (M/L)
$Pou6f2^{+}$ V1
$Pou6f2^{+};Nr5a2^{+}$
$Pou6f2^{+};Lmo3^{+}$
$Pou6f2^{+};Nr5a2^{+}$
$Pou6f2^{+};Lmo3^{+}$
F
$V1^{\text{Dual TF}}$ (D/V)
$Prdm8^{+}$ V1
$Prdm8^{+};Sp8^{+}$
$Prdm8^{+};FoxP4^{+}$
$Prdm8^{+};Sp8^{+}$
$Prdm8^{+};FoxP4^{+}$
G
V1 (R/C)
$V1^{\text{Pou6f2}}$ (R/C)
$V1^{\text{Sp8}}$ (R/C)
L3 vs L5
L3 vs L5
L3 vs L5
H
Inter-animal variation
SP8
6F2
5A2
6F2
LM3
3B3
PX1
3B2
5A2
FP4
MFA
I
Spatial Segregation (Summary)
SP8
6F2
5A2
6F2
LM3
3B3
PX1
3B2
5A2
FP4
MFA

**Figure 3.2 (preceding page). Spatial Segregation of V1 Interneuron Subpopulations**
(a) V1 interneurons in p0 L3-L5 segments of En1.nLacZ mice. D/V axis range: 132 to -265 µm; M/L axis range: 127 to 487 µm, 5th-95th percentiles from central canal. Contours represent density at the 30th-90th percentiles.

(b) Spatial clustering of V1Pou6f2/Nr5a2 interneurons (blue, Fa = 0.236) ($p < 0.00001$, one-tailed Monte Carlo test compared to parental V1).

(c) M/L biases in distributions of V1Sp8 (Xepicenter = 162 µm) and V1Pou6f2 (Xepi = 403 µm) interneurons. p < 1 x 10-20, Wilcoxon Rank Sum test in x-axis, V1Sp8 or V1Pou6f2 vs V1Parental, and V1Sp8 vs V1Pou6f2.

(d) D/V biases in distributions of V1Pou6f2 (Yepi = 66 µm), V1FoxP4 (Yepi = -158 µm), and V1MafA (Yepi = -277 µm) interneurons. V1Sp8 interneurons (Yepi = 72 µm) also occupy a dorsal position. p < 1 x 10-20, Wilcoxon Rank Sum test in y-axis for V1Pou6f2, V1FoxP4, V1MafA, or V1Sp8 vs V1Parental.

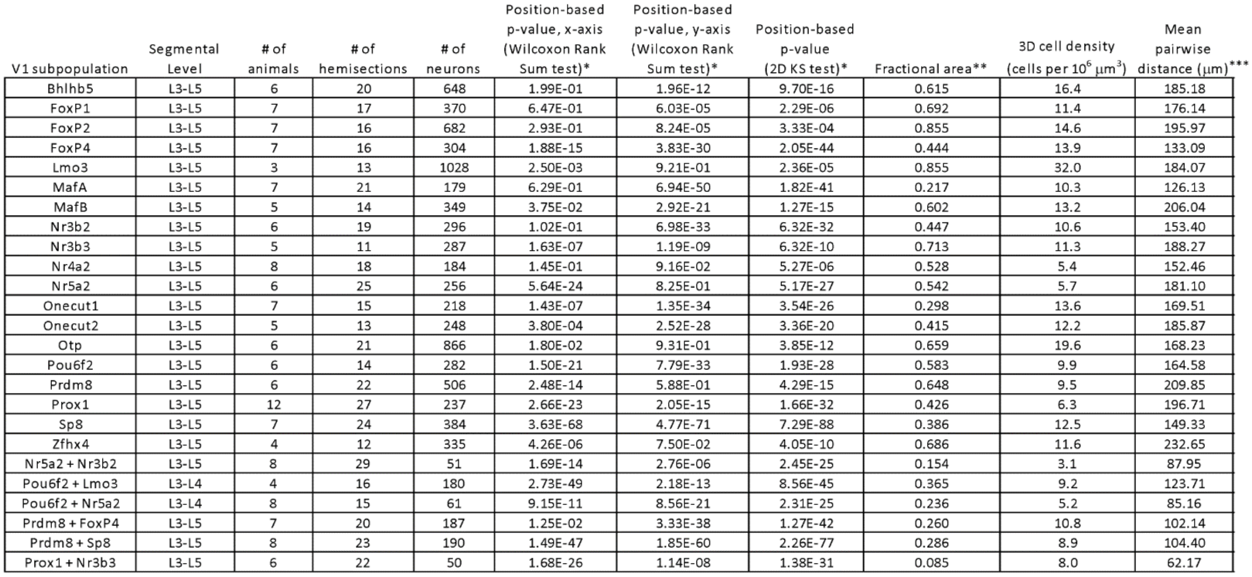
(e) Subdivision of V1Pou6f2 interneurons into medial (Nr5a2+, blue) and lateral (Lmo3+, red) subsets in p0 L3-L4 spinal segments.

(f) V1Prdm8 interneurons fractionate into dorsal Sp8+ (blue) and ventral FoxP4+ (red) composite groups.

(g) V1, V1Pou6f2, and V1Sp8 settling position at L3 (blue) or L5 (red) in p0 mice. $p < 0.0001$ for L3 vs L5, 2D KS test.

(h) Constancy of x,y position (mean ± SD) for V1 interneurons expressing Sp8 (n = 7), Pou6f2/Nr5a2 (n = 8), Pou6f2/Lmo3 (n = 4), FoxP4 (n = 7), Nr3b2/Nr5a2 (n = 8), Nr3b3/Prox1 (n = 6), and MafA (n = 7 animals).

(i) Spatial distributions of seven V1 subsets. Contours represent 60th-90th percentile densities.



| V1 subpopulation | Segmental Level | # of animals | # of hemisections | # of neurons | Position-based p-value, x-axis (Wilcoxon Rank Sum test)* | Position-based p-value, y-axis (Wilcoxon Rank Sum test)* | Position-based p-value (2D KS test)* | Fractional area** | 3D cell density (cells per $10^6$ $\mu m^3$) | Mean pairwise distance ($\mu m$)*** |
|---|---|---|---|---|---|---|---|---|---|---|
| Bhlhb5 | L3-L5 | 6 | 20 | 648 | 1.99E-01 | 1.96E-12 | 9.70E-16 | 0.615 | 16.4 | 185.18 |
| FoxP1 | L3-L5 | 7 | 17 | 370 | 6.47E-01 | 6.03E-05 | 2.29E-06 | 0.692 | 11.4 | 176.14 |
| FoxP2 | L3-L5 | 7 | 16 | 682 | 2.93E-01 | 8.24E-05 | 3.33E-04 | 0.855 | 14.6 | 195.97 |
| FoxP4 | L3-L5 | 7 | 16 | 304 | 1.88E-15 | 3.83E-30 | 2.05E-44 | 0.444 | 13.9 | 133.09 |
| Lmo3 | L3-L5 | 3 | 13 | 1028 | 2.50E-03 | 9.21E-01 | 2.36E-05 | 0.855 | 32.0 | 184.07 |
| MafA | L3-L5 | 7 | 21 | 179 | 6.29E-01 | 6.94E-50 | 1.82E-41 | 0.217 | 10.3 | 126.13 |
| MafB | L3-L5 | 5 | 14 | 349 | 3.75E-02 | 2.92E-21 | 1.27E-15 | 0.602 | 13.2 | 206.04 |
| Nr3b2 | L3-L5 | 6 | 19 | 296 | 1.02E-01 | 6.98E-33 | 6.32E-32 | 0.447 | 10.6 | 153.40 |
| Nr3b3 | L3-L5 | 5 | 11 | 287 | 1.63E-07 | 1.19E-09 | 6.32E-10 | 0.713 | 11.3 | 188.27 |
| Nr4a2 | L3-L5 | 8 | 18 | 184 | 1.45E-01 | 9.16E-02 | 5.27E-06 | 0.528 | 5.4 | 152.46 |
| Nr5a2 | L3-L5 | 6 | 25 | 256 | 5.64E-24 | 8.25E-01 | 5.17E-27 | 0.542 | 5.7 | 181.10 |
| Onecut1 | L3-L5 | 7 | 15 | 218 | 1.43E-07 | 1.35E-34 | 3.54E-26 | 0.298 | 13.6 | 169.51 |
| Onecut2 | L3-L5 | 5 | 13 | 248 | 3.80E-04 | 2.52E-28 | 3.36E-20 | 0.415 | 12.2 | 185.87 |
| Otp | L3-L5 | 6 | 21 | 866 | 1.80E-02 | 9.31E-01 | 3.85E-12 | 0.659 | 19.6 | 168.23 |
| Pou6f2 | L3-L5 | 6 | 14 | 282 | 1.50E-21 | 7.79E-33 | 1.93E-28 | 0.583 | 9.9 | 164.58 |
| Prdm8 | L3-L5 | 6 | 22 | 506 | 2.48E-14 | 5.88E-01 | 4.29E-15 | 0.648 | 9.5 | 209.85 |
| Prox1 | L3-L5 | 12 | 27 | 237 | 2.66E-23 | 2.05E-15 | 1.66E-32 | 0.426 | 6.3 | 196.71 |
| Sp8 | L3-L5 | 7 | 24 | 384 | 3.63E-68 | 4.77E-71 | 7.29E-88 | 0.386 | 12.5 | 149.33 |
| Zfhx4 | L3-L5 | 4 | 12 | 335 | 4.26E-06 | 7.50E-02 | 4.05E-10 | 0.686 | 11.6 | 232.65 |
| Nr5a2 + Nr3b2 | L3-L5 | 8 | 29 | 51 | 1.69E-14 | 2.76E-06 | 2.45E-25 | 0.154 | 3.1 | 87.95 |
| Pou6f2 + Lmo3 | L3-L4 | 4 | 16 | 180 | 2.73E-49 | 2.18E-13 | 8.56E-45 | 0.365 | 9.2 | 123.71 |
| Pou6f2 + Nr5a2 | L3-L4 | 8 | 15 | 61 | 9.15E-11 | 8.56E-21 | 2.31E-25 | 0.236 | 5.2 | 85.16 |
| Prdm8 + FoxP4 | L3-L5 | 7 | 20 | 187 | 1.25E-02 | 3.33E-38 | 1.27E-42 | 0.260 | 10.8 | 102.14 |
| Prdm8 + Sp8 | L3-L5 | 8 | 23 | 190 | 1.49E-47 | 1.85E-60 | 2.26E-77 | 0.286 | 8.9 | 104.40 |
| Prox1 + Nr3b3 | L3-L5 | 6 | 22 | 50 | 1.68E-26 | 1.14E-08 | 1.38E-31 | 0.085 | 8.0 | 62.17 |

**Table 3.1. Statistical Analysis of Spatial Distributions of V1 Subpopulations.**
Summary of spatial metrics and statistical analysis for each of the V1 subpopulations. *p-values correspond to comparisons of the distributions of a given V1 subpopulation and the parental V1 populations. ** All V1 subpopulations covered a significant smaller area than the parental V1 population. (p<0.001 by one-tailed Monte Carlo test). *** The mean pairwise distance for parental V1 interneurons is 212.09 um.

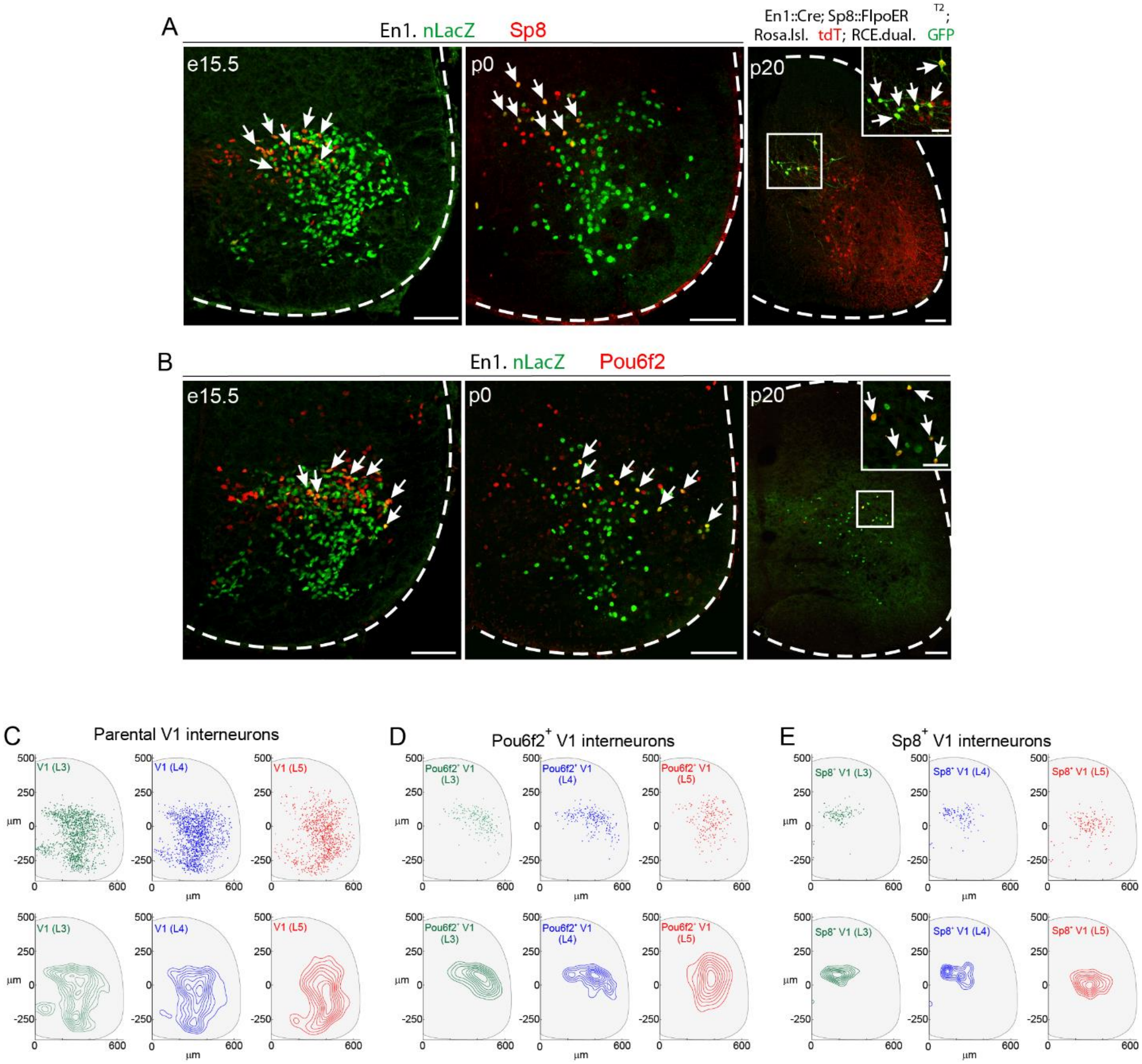

A
En1. nLacZ Sp8
En1::Cre; Sp8::FlpoER T2;
Rosa.lsl. tdT; RCE.dual. GFP
e15.5
p0
p20
B
En1. nLacZ Pou6f2
e15.5
p0
p20
C
Parental V1 interneurons
V1 (L3)
V1 (L4)
V1 (L5)
D
Pou6f2+ V1 interneurons
Pou6f2+ V1 (L3)
Pou6f2+ V1 (L4)
Pou6f2+ V1 (L5)
E
Sp8+ V1 interneurons
Sp8+ V1 (L3)
Sp8+ V1 (L4)
Sp8+ V1 (L5)
500
250
0
-250
µm
0
600
µm

**Figure 3.3 (preceding page). Constancy of V1Sp8 and V1Pou6f2 Interneuron Position, and Analysis of Rostrocaudal Spatial Distributions**

(a-b) Lumbar spinal cords from e15.5, p0, or p20 En1.nLacZ or En1::Cre; Sp8::FlpoERT2; Rosa.lsl.tdT; RCE.dual.GFP mice show similar dorsomedial locations for V1Sp8 interneurons (arrows, A) and similar dorsolateral locations for V1Pou6f2 interneurons (arrows, B), independent of age. Scale bars = 100µm or 50 µm (inset).

(c) V1 interneuron distributions at single segmental lumbar spinal levels (L3: n = 3 animals, 11 hemisections, 1332 cells; L4: n = 3 animals, 13 hemisections, 1541 cells; L5: n = 3 animals, 8 hemisections, 866 cells). While each segmental level showed a statistically significant difference in spatial distribution, L5 varied most among the three levels (p < 10-30 for L3 or L4 vs L5; p < 10-4 for L3 vs L4, 2D KolmogorovSmirnov test).

(d) Analysis of V1Pou6f2 interneurons at single segmental lumbar spinal levels (L3: n = 8 animals, 12 hemisections, 182 cells; L4: n = 7 animals, 14 hemisections, 227 cells; L5: n = 7 animals, 12 hemisections, n = 204 cells). L3 and L4 distributions were not significantly different (p = 0.10, 2D KS test), whereas L5 differed from both L3 and L4 (p < 0.001, 2D KS test).

(e) Analysis of V1Sp8 interneurons at single segmental lumbar spinal levels (L3: n = 6 animals, 9 hemisections, 122 cells; L4: n = 6 animals, 9 hemisections, 124 cells; L5: n = 6 animals, 9 hemisections, 146 cells). Similar to the parental V1 population, L5 varied most among the three levels (p < 10-14 for L3 or L4 vs L5; p = 0.023 for L3 vs L4, 2D KS test).

### 3.2.2 Incorporating spatial information into our computational analysis.

The aforementioned localization of spinal interneurons in stereotyped spatial domains, prompted us to ask whether the incorporation of spatial information can refine estimates of V1 group diversity. For this spatial analysis, we divided the ventral spinal cord into 196 bins and defined cell-type fractions for each bin. Similarly, we divided the spatial expression data into bins, and generalized the Bayesian analysis described in the previous section to model these spatially-resolved data.

Mathematically, the inclusion of spatial information translates into an expansion in each measurement index to accommodate spatial location, $M^1_{\mathrm{i,\,x}}$. Likewise, fractional values can now be inferred in each location of the grid rewriting $f_k$ as $f_{k,\,x}$ and, interpreting $f_k$ as resulting from spatial aggregation, $f_k = \sum_{\mathrm{x}} f_{k,\,x}$, where the sum is over a discrete grid covering the observed region. Rewriting equations 2.1 but now including spatial coordinates we obtain:

$$M^1_{\mathrm{a,\,x}} = \Sigma_{\mathrm{k}}\, f_{k,x}\, J_{k,a} + \varepsilon_{\mathrm{a,x}} \qquad \varepsilon_{\mathrm{a,x}} \sim \mathrm{N}(0,\sigma^2_{\mathrm{a,x}}) \qquad \mathrm{a} = 1,\ldots,19$$

$$M^2_{\mathrm{a,\,b}} = \Sigma_{\mathrm{k,x}}\, f_{k,x}\, J_{k,a}\, J_{k,b} + \varepsilon_{\mathrm{ab}} \qquad \varepsilon_{\mathrm{ab}} \sim \mathrm{N}(0,\sigma^2_{\mathrm{ab}}) \qquad \mathrm{a,\,b} = 1,\ldots,19 \tag{3.1}$$

In our formulation, we only expand measurement performed on single transcription factor spatial distributions. Equations involving pairs of transcription factors retain their global dependencies (collapsing the spatial component by summing out this dependency) because these spatial distributions were not measured.

In practice, the locations of cells expressing transcription factor *a* in each experiment correspond to sets of two-dimensional coordinates $p_{a,n}$, which we treat as samples from a smooth probability distribution. To infer the latter, we express it as a linear combination of Gaussian smoothing kernels centered on the observed points $p_{a,n}$ [Wasserman, 2006]. Moreover, we are interested in population averages, so we further average the smoothed density over all the experiments. The spatial density $\mathrm{M}^1_{\mathrm{a,\,x}}$ is therefore obtained as:

$$\mathrm{M}^1_{\mathrm{a,\,x}} = \gamma_{\mathrm{a}} \;\langle\, \textstyle\sum_{\mathrm{n}} \mathrm{h}(p_{a,n} - \mathrm{x}) \,\rangle, \tag{3.2}$$

subject to:

$$\textstyle\sum_{\mathrm{x}} \mathrm{M}^1_{\mathrm{a,\,x}} = \mathrm{M}^1_{\mathrm{a}} \tag{3.3}$$

where h is the Gaussian smoothing kernel and the average ⟨·⟩ is over repeated experiments. Note that in practice we only evaluate the continuous density in the discrete grid of the x's (dividing space into a 14x14 grid, 196 locations), the width of the Gaussian kernel was considered to be half the distance between grid points, and the normalization factor $\gamma_{\mathrm{i}}$ enforces constraint 3.3.

Similarly to the elimination of possible cell types by absence of co-expression, $\mathrm{M}^2_{\mathrm{a,\,b}} = 0$, consider for each transcription factor *a,* the locations x with $\mathrm{M}^1_{\mathrm{a,\,x}} = 0$. It follows from (3.1) that for all those k's with $J_{k,i} = 1$ we can set $f_{k,x} = 0$ and eliminate the equation for $\mathrm{M}^1_{\mathrm{a,\,x}}$. Although these constraints do not eliminate any cell type here, we find that the number of equations for $\mathrm{M}^1_{\mathrm{a,\,x}}$ can be reduced by half, improving the computational efficiency of the inference methods discussed below. For each k, let us call $\mathrm{X}_{\mathrm{k}}$ the set of locations x in which we have not set $f_{k,x} = 0$. Again, we can divide equations (3.1) by the noise standard deviations and group them as a linear regression problem, where now the observed vector *Y* represents the measured data $\mathrm{M}^1_{\mathrm{a,\,x}}$ and $\mathrm{M}^2_{\mathrm{a,\,b}}$, normalized by the corresponding $\sigma_{a,x}$ and $\sigma_{ij}$. As in chapter 2, the solutions $f_{k,x}$ are constrained by $f_{k,x} \geq 0$, $\sum_k \sum_{x \in Xk} f_{k,x} = 1$ and, to identify solutions predicting non-zero measurements, we impose on $f_{k,x}$ constraints similar to (2.4).

*Extending NNCLS methodology to accommodate spatial information*

When incorporating spatial information, the optimization problem is of exactly the same form as in chapter 2, equations (2.6) - (2.7). However, we expand the fractional values as described above modelling $f_{k,x}$ in the vector $f$ and the corresponding spatial observations in $Y$. To study the degeneracy of the solutions we calculate $f^{min}_k$ and $f^{max}_k$ (Figure 3.4). In this case, the inclusion of spatial information tightens the bounds between $f^{min}_k$ and $f^{max}_k$ (compare against Figure 2.3). Furthermore, many of the $f^{min}_k$ values are bounded away from zero, indicating that these cell types are essential for achieving the minimal error. Nonetheless, according to this method we would need more than 250 cell types to explain the data, contradicting our prior belief that a small cell type subset comprises V1.

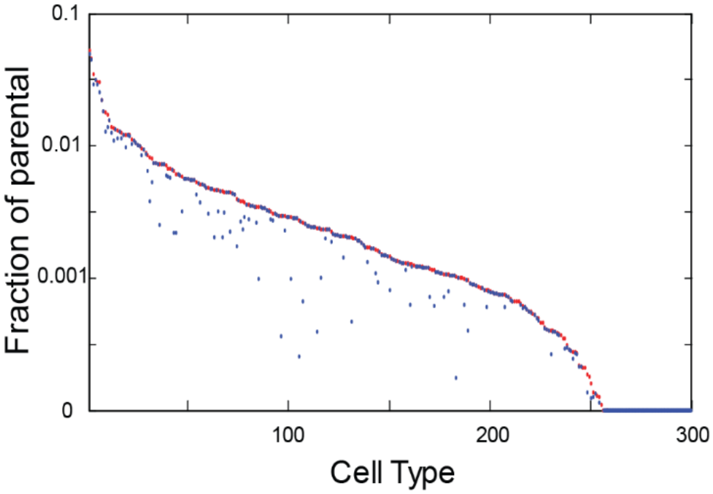


**Figure 3.4. Bounds on the fractional values achieved by the NNCLS solution incorporating spatial information.**

Inclusion of spatial information tightens the bounds on the minimum and maximum values that each fractional value can achieve (compared against Figure 2.3). $f^{min}_k$ and $f^{max}_k$ are indicated by blue and red dots respectively.

*Inclusion of Spatial Information into our Bayesian formulation*

The inclusion of spatial information into our Bayesian formulation transforms the prior, adapting it for regression problems that are sparse at the group level. This in turn, guarantees that all of the spatial coordinates for a given cell type k are turned on or off together. To achieve this goal, we rewrite the prior as:

$$\mathrm{p}(b_k|\, a) = \mathrm{a}^{bk}\,(1{-}\mathrm{a})^{1-bk}$$
$$\mathrm{p}(\mathrm{z}_{k,x}|\, \tau^2) = \mathrm{N}(0, \tau^2) \tag{3.4}$$

such that

$$f_{k,x} = b_k\, z_{k,x} \tag{3.5}$$

This formulation is similar to [Hernandez-Lobato, 2013], but we perform inference by sampling instead of using the expectation maximization algorithm. Inference is achieved as in chapter 2 but this time, constraints are modified as:

$$z_{k,x} \geq 0, \qquad \textstyle\sum_{\mathrm{k}} \sum_{\mathrm{x}\in\mathrm{Xk}} b_k\, z_{k,x} \leq 1, \qquad \sum_{\mathrm{k}} \sum_{\mathrm{x}\in\mathrm{Xk}} q_{i,k,x}\, b_k\, z_{k,x} > 0$$

In practice, to calculate the Bayesian spatial solution, for computational ease, the 256 most relevant cell types obtained by the non-spatial solution are used to pre-select the possible cell types (non-spatial posterior inclusion probability greater than 0.01). Using this information, our estimates of cell diversity are refined yielding inference on the spatial location of the estimated cell types.

### *Computational Validation Experiments*

Before applying our inference algorithm to real data, we performed computational experiments, echoing the procedure in chapter 2. In this case, we repeated the leave-one-out cross-validation experiment, this time incorporating spatial information. Again, most of the cross-validated cell-type fractions fall within two standard deviations from the measured fractions (Figure 3.5a). Inferred spatial distributions from the cross-validation experiment possess similar epicenters as their "true" counterparts, however, they are somewhat broadened (Figure 3.5b-c); this trend is consistent with the fact that slightly fewer data points are used to constrain the cross-validated experiment compared to the full estimate.

To clarify the difference between experiments in Chapter 2 and 3, to perform leave-one-out cross-validation experiments not including spatial information, each of the measured fractions (single and paired transcription factors) was excluded from the measurements dataset and their values are then inferred. To perform leave-one-out cross-validation experiments including spatial information, each entire single transcription factor spatial distribution is removed in each experiment. Subsequently, removed spatial distributions are inferred and the total fraction of cells is calculated adding the fraction of cells at each spatial location.

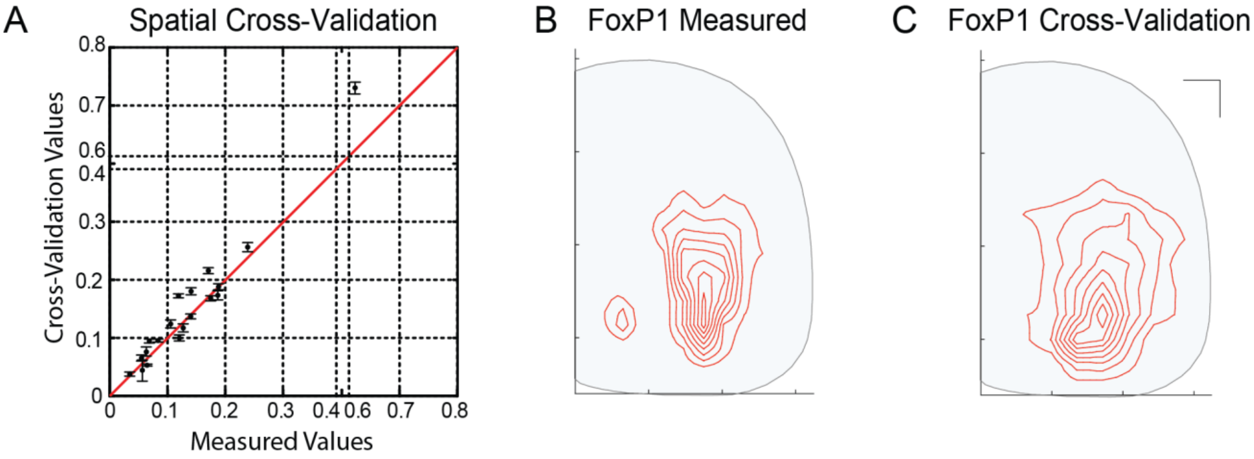


**Figure 3.5: Computational validation suggests that inference algorithms are robust and effective.**

(a) Leave-one-out cross validation experiment showing good agreement between measured and cross-validated values in the spatial algorithm.

(b), (c) Spatial distribution of FoxP1 cells and inferred spatial distribution after performing a cross-validation experiment in which all the spatial information about FoxP1 cells is removed. The recovered spatial distributions are broadened slightly; this trend is observed consistently in our spatial cross-validation experiments (other experiments not shown), and is consistent with the fact that slightly fewer data points are used to constrain the cross-validated experiment compared to the full estimate.

### 3.2.3 Spatial information reveals further V1 interneuron diversity

We next sought out to compute estimates of cell type diversity within V1 interneurons considering three sets of data: (i) the fraction of neurons within the parental V1 population that express each of the 19 TFs (Figure 2.2a), (ii) the fractions of neurons co-expressing various pairs of TFs (Figure 2.2b-c), and (iii) the position of V1 interneurons expressing each of the 19 TFs (Figure 3.1a).

Incorporating spatial information into the Bayesian analysis increased the number of cell types that the HMC sampler selected per iteration, as well as the degree of confidence in the inferred expression profiles.  The number of selected cell types per iteration increased from about 30 in the non-spatial setting to 50 ± 2 (mean ± SD over all HMC samples; Figure 3.6a). And with the additional of spatial information just 75 total cell types are assigned posterior inclusion probabilities greater than 0.05, compared to 131 in the non-spatial setting. Moreover, many of these spatially-informed inclusion probabilities are now much closer to one (Figure 3.6E). We repeated the grouping procedure for these 75 total cell types and uncovered 57 candidate expression profiles, most identified with very high inclusion probabilities (Figure 3.6b) and significantly reduced ambiguities in their expression profiles compared to the non-spatial results. Comparison of the inclusion probabilities obtained before and after incorporation of spatial information emphasizes a much more confident assignment of cell types (Figure 3.6c). Thus we can now assign specific expression profiles to a majority of the approximately 50 predicted cell types, and can provide strong probabilistic constraints on the expression patterns of the remaining types.

An additional benefit of this spatial analysis is that it provides estimates of how each inferred cell type localizes in the ventral spinal cord (Figure 3.7). Many of the inferred cell types are localized in relatively compact, contiguous domains, covering the full positional spectrum of the parental V1 interneuron distribution, along both the dorso-ventral and medio-lateral axes. Notably, one inferred cell type with the expression profile of Renshaw interneurons (expressing MafB, Onecut1 and Onecut2) is predicted to be confined to the most ventral region within the parental V1 population (Figure 3.7i), in agreement with their known settling position [Alvarez and Fyffe, 2007]; [Stam et al, 2012]. Other inferred cell types, characterized by FoxP2, FoxP4, Nr3b3, and/or Nr4a2 expression, showed clustered distributions ventral to the central canal, and dorsomedial to motor neurons (Figure 3.8).  Such distributions are similar to the proposed location of group Ia reciprocal interneurons [Hultborn, 1971], a subset of which are known to reside within the parental V1 population [Zhang et al, 2014]. Thus these findings document novel molecular and spatial diversity in the V1 interneuron population.

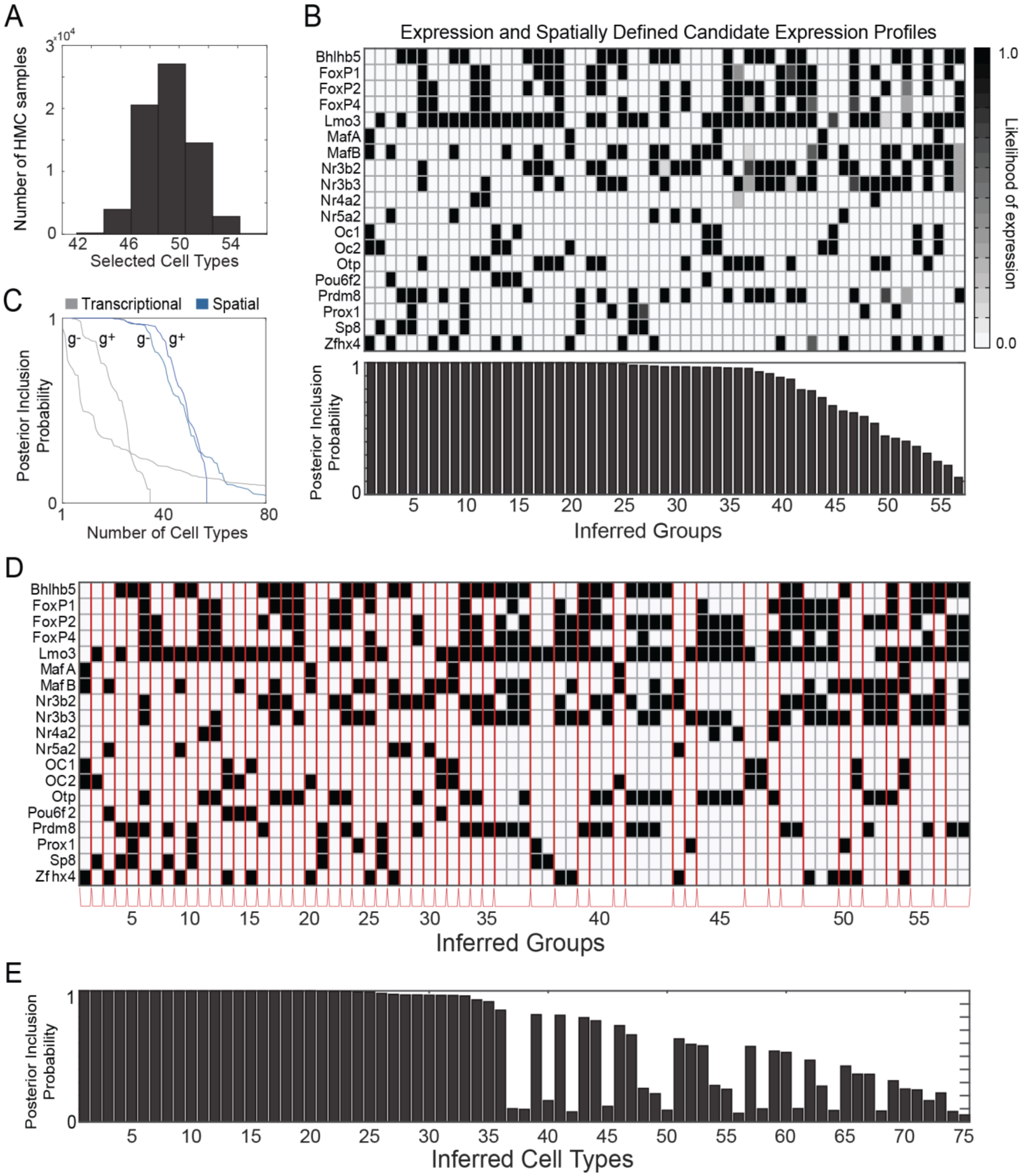

A
3 ×10^4
Number of HMC samples
0
1
2
42
46
50
54
Selected Cell Types
B
Expression and Spatially Defined Candidate Expression Profiles
Bhlhb5
FoxP1
FoxP2
FoxP4
Lmo3
MafA
MafB
Nr3b2
Nr3b3
Nr4a2
Nr5a2
Oc1
Oc2
Otp
Pou6f2
Prdm8
Prox1
Sp8
Zfhx4
1.0
0.0
Likelihood of expression
Posterior Inclusion Probability
1
0
5
10
15
20
25
30
35
40
45
50
55
Inferred Groups
C
Transcriptional
Spatial
g-
g+
g-
g+
Posterior Inclusion Probability
1
0
1
40
80
Number of Cell Types
D
Bhlhb5
FoxP1
FoxP2
FoxP4
Lmo3
MafA
MafB
Nr3b2
Nr3b3
Nr4a2
Nr5a2
OC1
OC2
Otp
Pou6f2
Prdm8
Prox1
Sp8
Zfhx4
5
10
15
20
25
30
35
40
45
50
55
Inferred Groups
E
Posterior Inclusion Probability
1
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
Inferred Cell Types

**Figure 3.6 (preceding page): Cell Types Revealed by Incorporating Transcription Factor Spatial Information.**

(a) Number of selected cell types in each HMC iteration.

(b) Condensed representation of candidate expression profiles of 57 V1 groups. Gray scale indicates the likelihood that each TF is expressed within the group. Bottom, posterior inclusion probability for each V1 group.

(c) Posterior inclusion probability for expression-inferred cell types and groups (gray), and expression-and-spatially-inferred cell types and groups (blue); “g+” indicates groups, and “g-” cell types

(d) Transcription factors expressed by transcriptionally and spatially defined cell types with a posterior inclusion probability greater than 5 percent. Group members are arranged in between red lines.

(e) Posterior inclusion probability for cell types in (d). Cell types are ranked by group-level inclusion probability.

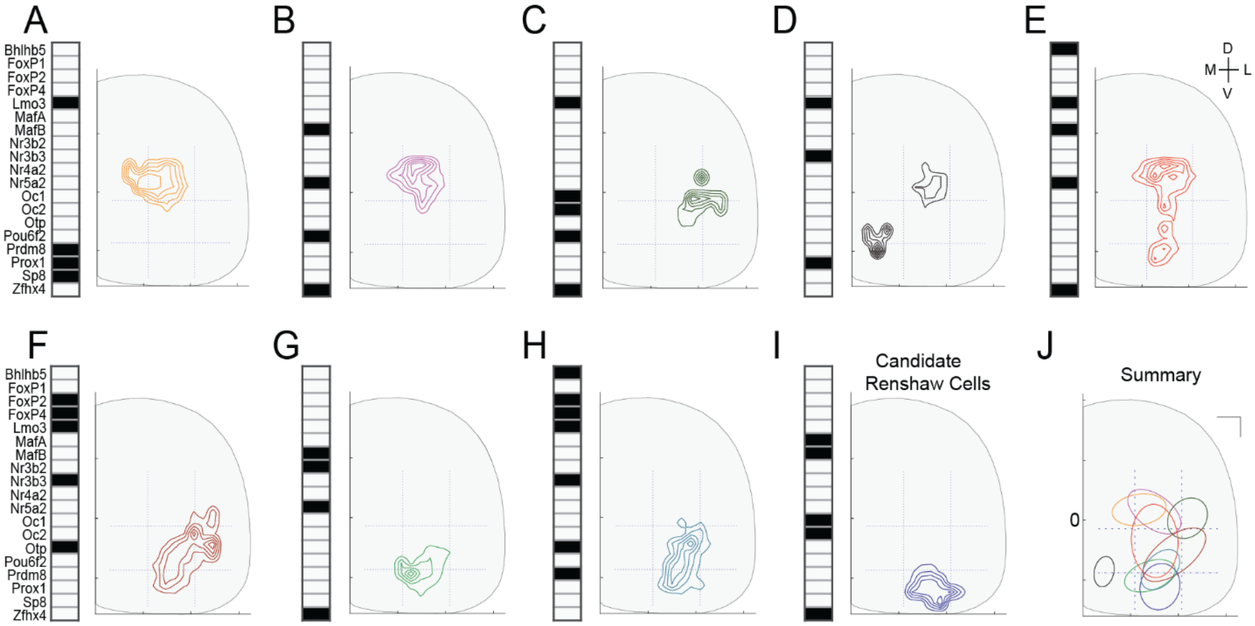


**Figure 3.7: Inferred cell type spatial distributions segregate V1 interneurons into compact domains.**
(a- i) Positional distributions of inferred V1 cell types. These populations are confined to compact spatial domains.

(i), spatial distribution of an inferred cell type corresponding to candidate Renshaw interneurons, defined by expression of known Renshaw markers (MafB, Oc1, and Oc2) and localization in an extreme ventral position.

(j) Spatial distributions from cell types in (A-I) aggregated in a single plot. Each cell type is represented by its confidence ellipse under a Gaussian approximation to the posterior spatial distribution of each cell type (66% confidence ellipse). Scale bar = 100 μm.

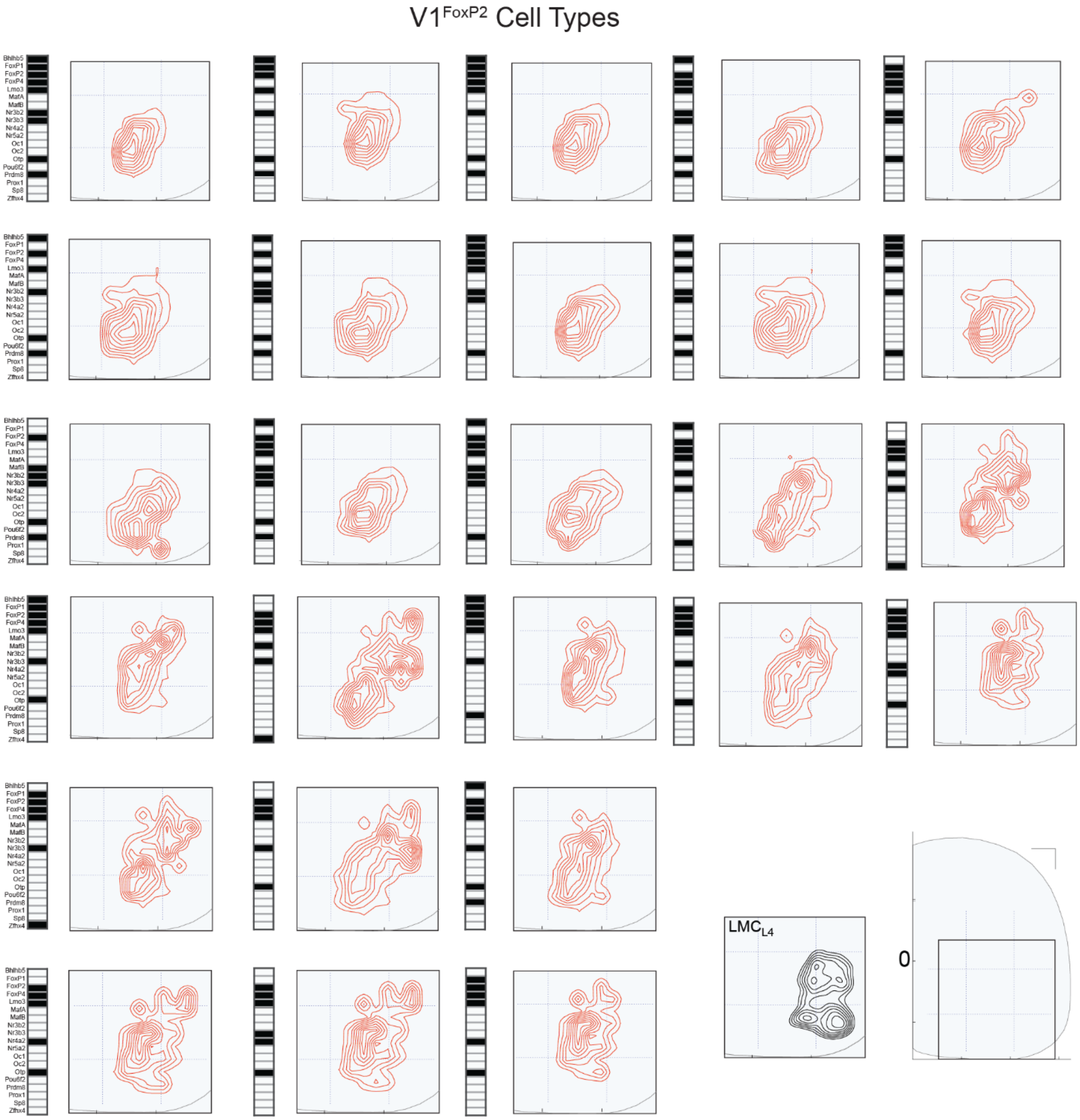


**Figure 3.8: Spatial Distributions of FoxP2-Expressing V1 Cell Types reveal highly overlapping cell types.** The spatial distributions of 26 inferred V1 cell types contained within the FoxP2 clade. Many of these distributions are clustered in a spatial domain ventral to the central canal (position 0), and dorsomedial to the putative motor neuron domain (MN). Such distributions occupy a similar relative position to that reported for group Ia reciprocal interneurons in the adult cat (Hultborn, 1971). Motor neurons in the lateral motor column are depicted in black. Combined with the prior suggestion that some group Ia reciprocal interneurons express FoxP2 (Benito-Gonzalez, 2012), these represent candidate cell types that may correspond to reciprocal inhibitory interneurons. Note that these cell types exhibit highly overlapping spatial domains, and therefore are distinguished solely based on molecular phenotype.

### 3.2.4 A cladistic analysis of transcription factor expression

Next we addressed the issue of the number of transcription factors needed to define a neuronal cell type. We find that 5 ± 2 (mean ± SD) of 19 considered transcription factors are expressed within each cell type (Figure 3.6b). To characterize the minimal number of transcription factors that would be needed to provide selective access to an individual cell type, we developed a classification scheme that relies on a recursive algorithm to sequentially subdivide the parental population, and arrange every cell type along a clade diagram. In this representation, the central node of the diagram corresponds to the full V1 population, with branches representing transcription factors expressed in a mutually exclusive fashion and covering the highest fraction of the parental population. This process is repeated until the cell types from which the analysis is constructed are revealed at the extremities of the plot. Additional nodes (represented as circles) are associated with remaining cell types that do not express mutually exclusive transcription factors. In cases in which there are no mutually exclusive transcription factors, branches combine transcription factors representing the highest fraction of cells in the node. Finally, nodes depicted with a bar on top of the name of a transcription factor represent cell types that are defined by the absence of transcription factors at the node. These nodes would require a repressor strategy to be targeted (similar to the Gal80 repressor in Drosophila Melanogaster). See Algorithm 3.1 for full details.

**Algorithm 3** Clade Diagram

```
Denote CT: Set of cell types.
Denote CTf: Fraction of cells expressed by each cell type in CT.
Denote TF: Set of transcription factors expressed by members of CT.
Denote $\mathscr{M}$: Set of mutually exclusive transcription factors.
BEGIN
Input Data: TF,CT, CTf.
Compute $\mathscr{M}$ from CT, CTf maximing fraction of cells.
for Each member $\mathscr{M}_i$ in $\mathscr{M}$ do
    Create a node called $\mathscr{M}_i$
    Create TFn, CTn CTfn removing cell types expressing $\mathscr{M}_i$ from (TF,CT,CTf)
    Remove $\mathscr{M}_i$ from TFn,CTn,CTfn.
    if Ctn contains an empty cell type then
        Create an empty Node
        Remove Cell Type from CTn.
    end if
    if CTn is not empty then
              Run Clade Diagram, input data: TFn,CTn,CTfn.
    end if
end for

if $\mathscr{M}$ is empty but TF is not then
    Compute $\mathscr{C}$ from TF, set of Transcription factors that cover most of the population.
    for Each member $\mathscr{C}_i$ in $\mathscr{C}$ do
        Create a node called $\mathscr{C}_i$
        Create TFn, CTn CTfn removing cell types expressing $\mathscr{C}_i$ from (TF,CT,CTf)
        Remove $\mathscr{C}_i$ from TFn,CTn,CTfn.
        if Ctn contains an empty cell type then
            Create an empty Node
            Remove Cell Type from CTn.
        end if
        if CTn is not empty then
                  Run Clade Diagram, input data: TFn,CTn,CTfn.
        end if
    end for
end if
Backtrack and remove empty nodes where no other nodes arises from the parent.
```

To help the reader visualize the construction of the graphical representation, we generated different clade diagrams (Figure 3.9) for cell types with increasing transcriptional complexity. Complexity is increased to portray how additional cell types diversify existing clades or create new ones. In addition, examples are included describing the incorporation of remaining (not mutually exclusive) cell types and cell types whose transcriptional profile is determined by the absence of certain transcription factors.

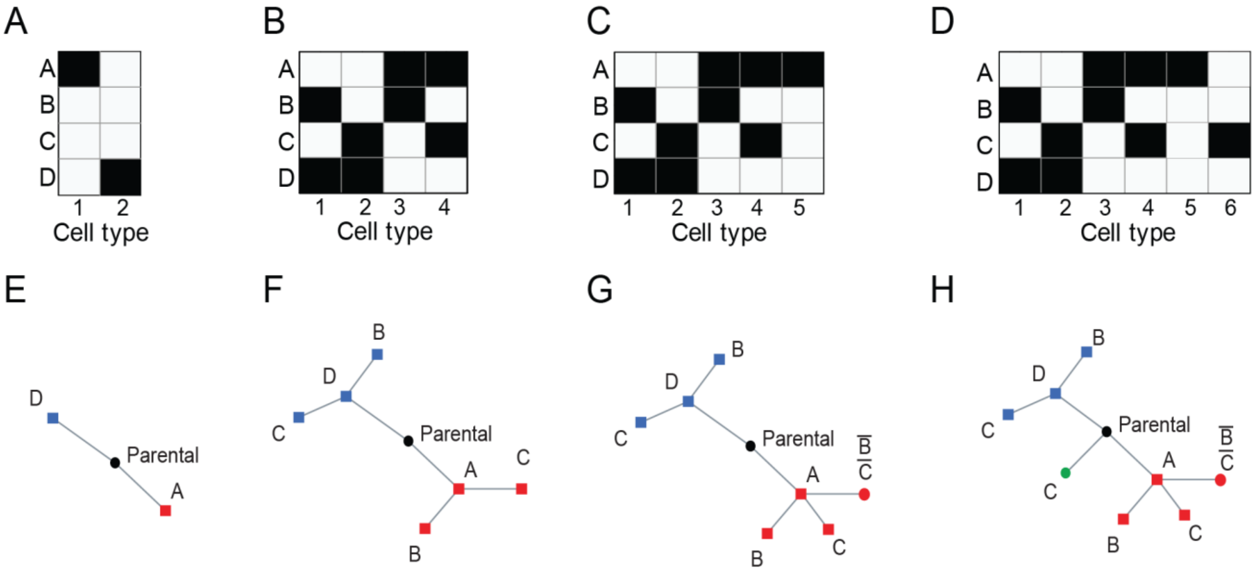


**Figure 3.9: Clade diagrams organize cell populations in a transcriptionally mutually exclusive fashion.**

Examples of clade diagrams built on cell types of increasing transcriptional complexity.

(a) Simplest representation composed of two mutually exclusive populations.

(b) Cell types in (a) are divided given two mutually exclusive transcription factors, termed b and c.

(c) Same as in (b). A cell type is added whose transcriptional profile is determined by the absence of the transcription factors at the node.

(d) Same as in (c). A cell type is added to highlight how remaining cell types (not mutually exclusive) are incorporated into the diagram.

(e) - (h) Clade diagrams corresponding to cell types in (a)-(d).

We sought to identify mutually exclusive transcription factors whose expression covers the highest fraction of the parental V1 population, excluding Lmo3, which itself is expressed in 74% of all V1 interneurons. This analysis reveals that 64% of the V1 parental population can be divided into four main clades on the basis of the mutually exclusive expression of FoxP2, Pou6f2, Sp8, or MafA (Figure 3.10a). Each clade contains from 2 to 17 cell types, with a total of 38 cell types falling within the 4 main clades (Figure 3.10b). The minimum number of transcription factors needed to target each group ranges from 2 (in the case of MafA, Zfhx4) to 6 (as in the final leaves of the FoxP2 clade) with a mean number of 4 ± 1 (mean ± SD).

Analysis of cladal settling position demonstrated that while the MafA, Sp8, and Pou6f2 clades exhibited little spatial overlap, the FoxP2 clade displayed a broad spatial distribution with significant overlap with the other three clades (Figure 3.10c). At the second tier of our clade analysis, V1 subclasses become more restricted spatially (Figure 3.10d), with additional spatial restrictions for higher tiers (not shown). Cell types within the Pou6f2 clade exhibit medio-lateral gradations in their spatial distributions, determined by the expression of the transcription factors Nr5a2 and Lmo3 respectively. Similarly, cell types within the Sp8 clade segregate by position, according to the presence or absence of Onecut2. In many cases, however, cell types within a single clade showed overlapping spatial distributions, exemplified by the FoxP2 clade, which is characterized by numerous intermingled cell types, with no statistically significant difference among their centroid coordinates (Figure 3.8). In these cases, inferred cell types can be distinguished solely by their transcriptional profiles. In summary, this cladistic analysis provides predictive insight into the relative contributions of individual transcription factors in delineating the hierarchy of V1 interneuron cell types.

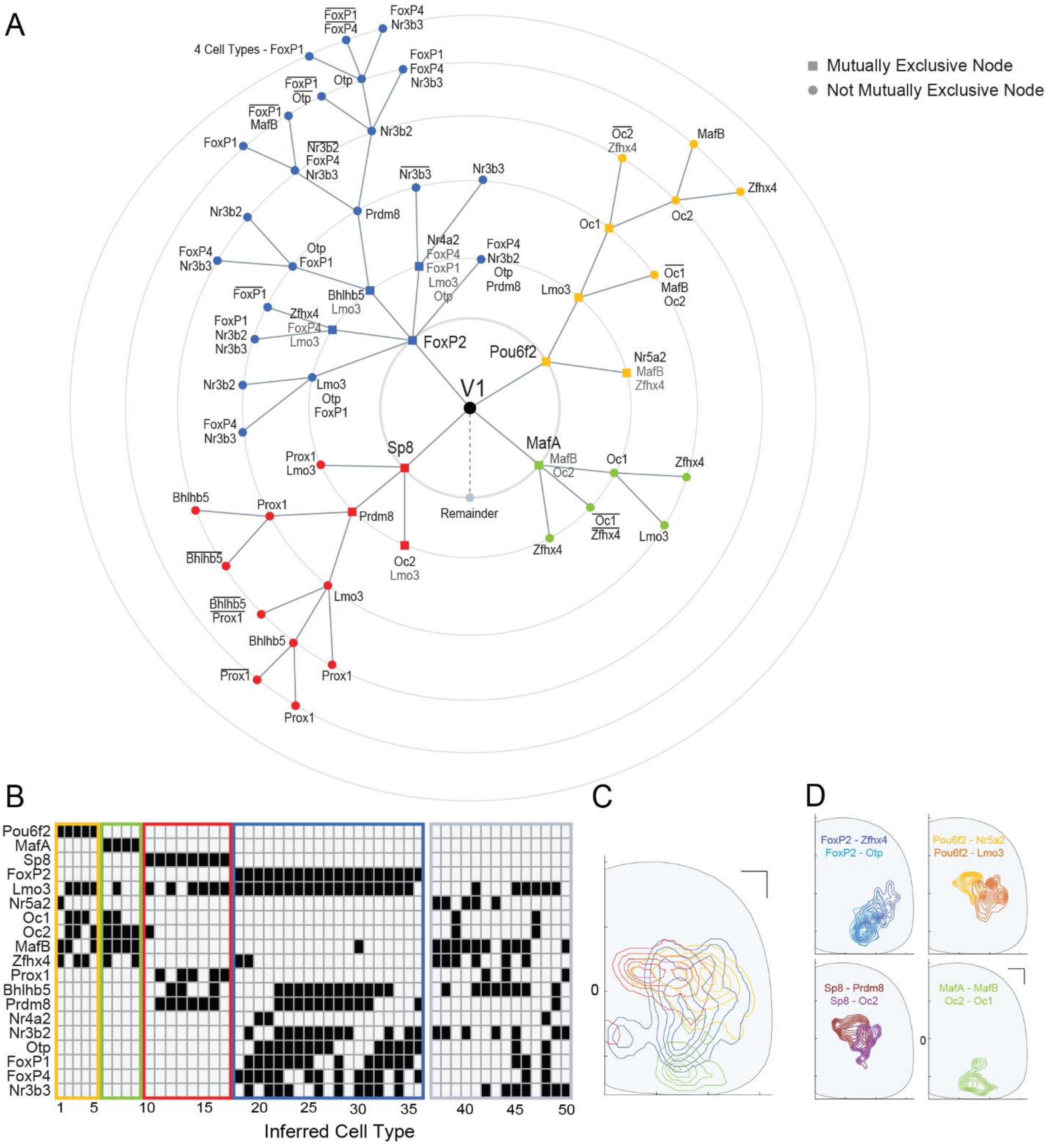

A
Mutually Exclusive Node
Not Mutually Exclusive Node
V1
Remainder
FoxP2
Pou6f2
Sp8
MafA
4 Cell Types - FoxP1
$\overline{FoxP1}$ $\overline{FoxP4}$
FoxP4 Nr3b3
Otp
FoxP1 FoxP4 Nr3b3
$\overline{FoxP1}$ $\overline{Otp}$
$\overline{FoxP1}$ MafB
Nr3b2
FoxP1
$\overline{Nr3b2}$ FoxP4 Nr3b3
$\overline{Nr3b3}$
Nr3b3
Prdm8
Nr3b2
Otp FoxP1
FoxP4 Nr3b3
Nr4a2 FoxP4 FoxP1 Lmo3 Otp
FoxP4 Nr3b2 Otp Prdm8
Bhlhb5 Lmo3
$\overline{FoxP1}$
Zfhx4 FoxP4 Lmo3
FoxP1 Nr3b2 Nr3b3
Nr3b2
Lmo3 Otp FoxP1
FoxP4 Nr3b3
$\overline{Oc2}$ Zfhx4
MafB
Zfhx4
Oc1
Oc2
Lmo3
$\overline{Oc1}$ MafB Oc2
Nr5a2 MafB Zfhx4
MafB Oc2
Oc1
Zfhx4
$\overline{Oc1}$ $\overline{Zfhx4}$
Zfhx4
Lmo3
Prox1 Lmo3
Bhlhb5
Prox1
Prdm8
Oc2 Lmo3
$\overline{Bhlhb5}$
Lmo3
$\overline{Bhlhb5}$ $\overline{Prox1}$
Bhlhb5
$\overline{Prox1}$
Prox1
Prox1
B
Pou6f2
MafA
Sp8
FoxP2
Lmo3
Nr5a2
Oc1
Oc2
MafB
Zfhx4
Prox1
Bhlhb5
Prdm8
Nr4a2
Nr3b2
Otp
FoxP1
FoxP4
Nr3b3
1
5
10
15
20
25
30
35
40
45
50
Inferred Cell Type
C
0
D
FoxP2 - Zfhx4
FoxP2 - Otp
Pou6f2 - Nr5a2
Pou6f2 - Lmo3
Sp8 - Prdm8
Sp8 - Oc2
MafA - MafB
Oc2 - Oc1
0

**Figure 3.10 (preceding page). Mutually Exclusive Cell Types Divide the V1 Parental Population into Four Clades.**

(a) Clade diagram constructed from the set of 50 cell types corresponding to the collection that occurs most frequently among the samples (mode of the posterior). Each terminal node corresponds to a cell type, with its TF profile obtained by traversing the diagram from the center to the outermost levels. Diagram is portrayed up to level 6 in the hierarchy and contains 15 cell types out of the 19 belonging to the FoxP2 clade. Bar above a TF name denotes lack of expression.

(b) Expression profiles of inferred cell types contained within each of the four clades. Gray box contains remaining cell types not expressed within $V1^{FoxP2}$, $V1^{MafA}$, $V1^{Pou6f2}$ or $V1^{Sp8}$ clades.

(c) $V1^{MafA}$, $V1^{Pou6f2}$ and $V1^{Sp8}$ clades represent mutually exclusive subsets but overlap spatially with the $V1^{FoxP2}$ clade. Scale bar =100 µm.

(d) V1 spatial distributions corresponding to subpopulations at the second tier of the clade diagram. Blue corresponds to cell types within the $V1^{FoxP2}$ clade, green correspond to $V1^{MafA}$, yellow corresponds to $V1^{Pou6f2}$ and red corresponds to $V1^{Sp8}$. Scale bar = 100 µm.

## 3.3 Discussion

Objective assessment of the extent of mammalian cellular diversity has remained challenging. The sparse Bayesian framework presented here provides a general method for characterizing cellular heterogeneity on the basis of sparsely sampled biological information. We have used this framework to study spinal V1 interneuron diversity. By analyzing the spatial expression densities of individual TFs as well as their patterns of pairwise expression, candidate expression profiles for ~50 inferred V1 cell types are provided. The integration of distinct phenotypic aspects of cellular heterogeneity, in this instance TF expression and settling position, markedly enhances the confidence in assignment of predicted V1 interneuron cell types. We note that this approach provides a general method for delineating the heterogeneity of cell types in any mixed tissue.

In this chapter, we have extended the Bayesian framework to perform group inference integrating spatial variables. In our analysis the inclusion of spatial data increased inferred cell type number by ~70%, and markedly enhanced confidence in the inferred expression profiles. This noticeable increment resolves the apparent indeterminacy allocated to the transcriptional only cell type group assignment, which permits one or more cell type to exist among the members of a group (see Chapter 2). Additionally, spatial information has a much stronger impact on cell type assignment than variation in the number of TFs, indicating that settling position carries significant cell type information which is independent from the information carried by the expression patterns of the 19 TFs examined here.

Indeterminacy it is not only portrayed on the posterior inclusion probability of each cell type, which is evidently reduced when using settling position along with transcriptional information, but also can be visualized in our clade diagram. This idea proposes another viewpoint when interpreting the clade diagram, one in which visualizing uncertainty is the goal of the diagram instead of mutually exclusiveness. This perspective reveals that uncertainty increases outwards from the center of the diagram, due to the increment in the number of transcription factors belonging to each cell type.

Finally, we note that transcriptional diversity could, in some instances, reflect variation in functional cell state, rather than indicating a distinct neuronal subtype. However, consistent with the genetic specification of cell type identity, we find that the position of transcriptionally distinct V1 subsets are segregated, stereotyped from animal to animal, and stable across development. The spatial segregation argues strongly against the ‘cell state’ possibility. Nevertheless, activity-shaped differences in Er81 expression in fast-spiking cortical interneurons have been shown to mark delay-type or non-delayed firing states [Dehorter et al, 2015]. In addition, activity-dependent induction of Npas4 expression has been described for cortical neurons [Lin et al., 2008], with implications for homeostatic regulation of sensitivity to inhibitory transmitters. Further studies will therefore be needed to dissect the functional consequences of V1 diversity, to resolve whether certain state-dependent functional properties are reflected in the diversity of V1 transcriptional profiles.

In summary, the results so far presented argue for a diverse cell type repertoire within V1 interneurons, localizing in compact spatial domains that, as we will see in chapter 5, impact on local micro circuitry connectivity. The purpose of the next chapter is to corroborate the validity of our statistical predictions and to assess the physiological impact of such transcriptional heterogeneity.

# Chapter 4

# Corroborating V1 diversity and cell type physiological properties

## 4.1 Introduction

The merits of our computational analysis depend critically on the ability to accurately infer cellular diversity. In chapters 2 and 3 we performed computational experiments aimed to demonstrate that our Bayesian approach can accurately infer cellular heterogeneity under idealized conditions in which a few cell types span the parental population and, ground truth is known a priori. Nonetheless, our computational validation leaves open the question of whether the inferred V1 cell types are actually present in vivo.

Accordingly, we sought to assess the biological accuracy of the Bayesian model's predictions by performing three additional experiments based on single cell quantitative PCR and further immunohistochemical measurements. These experiments aimed to validate different predictions of our analysis: expression profiles of cells, the prevalence of each profile within the parental population and, their settling position. Additionally, we explored supplementary computational experiments but this time based on biological correlates of cell type diversity in a different model organism and in the mouse cortex. In these cases, cell type heterogeneity has been previously extracted by computational clustering methodologies. These alternative experimental procedures will corroborate our Bayesian framework, validating out molecular and spatial predictions.

Lastly, we study whether molecular distinctions in V1 identity echo differences in interneuron physiology by performing electrophysiological experiments. Equipped with a molecular catalog represented by our clade diagram, we can direct the expression of a fluorescent reporter to visualize and recognize cells in different clades. By measuring the intrinsic physiological properties, we can explore if V1 interneurons can be distinguished by this additional characteristics. If V1 cell types turn out to be distinguished by molecular, morphological and physiological properties, this will represent a step further into identifying the true underlying diversity within V1. We can hypothesize that if a cell type fulfills a particular function within a neuronal circuit, evolution must have adapted aspects of its intrinsic properties to better serve that function.

## 4.2 Results

## 4.2.1 Validation of Bayesian model predictions

*Single-cell quantitative real-time PCR*

We first compared experimental findings from single-cell quantitative real-time PCR (qRT-PCR) data against inferred expression profiles from the Bayesian analysis (Figure 4.1). We focused on 15 transcription factors for which reliable qRT-PCR probes could be identified, and analyzed transcription factor expression within En1$^+$ neurons, isolated at random from p0 lumbar spinal segments of *En1::cre; RCE::lsl.GFP* mice.

First, we assessed whether qRT-PCR and immunocytochemistry gave comparable co-expression values. Appropriate thresholds for each gene were set, relative to the expression of the ubiquitously expressed gene β-actin, with the aim of comparing the measured qRT-PCR patterns against our immunohistochemical measurements. The threshold for each gene was calculated by minimizing the distance between a coexpression matrix calculated by qRT-PCR and immunohistochemistry. After applying these thresholds, transcription factors were classified either as 'expressed' or 'not expressed' within individual En1+ interneurons. The patterns of gene expression emerging from quantitative PCR exhibited high correlation with immunohistochemical data for both individual and paired transcription factor measurements (Figure 4.1a-c, correlation coefficient = 0.88).

Next, we asked whether qRT-PCR single-cell transcriptional patterns correlate with the clade results of our Bayesian analysis. Individual V1 interneuron gene expression profiles can be segregated along the four major inferred clade populations (Sp8, Pou6f2, MafA, and FoxP2), indicating that our computational predictions accurately reflect gene expression relationships in vivo (Figure 4.1d). Importantly, qRT-PCR identified closely related expression patterns predicted by the model. For example, V1Pou6f2+Lmo3 and V1Pou6f2+Nr5a2 subsets within the predicted Pou6f2 clade were revealed by qRT-PCR (Figure 4.1d). Although we do not possess enough coverage, cells that lack expression of any of the four cladal genes seem to follow the expression profiles of cell types belonging to the remainder of figure 3.10 (Figure 4.2). These results validate the general organization of the Bayesian cell-type predictions.

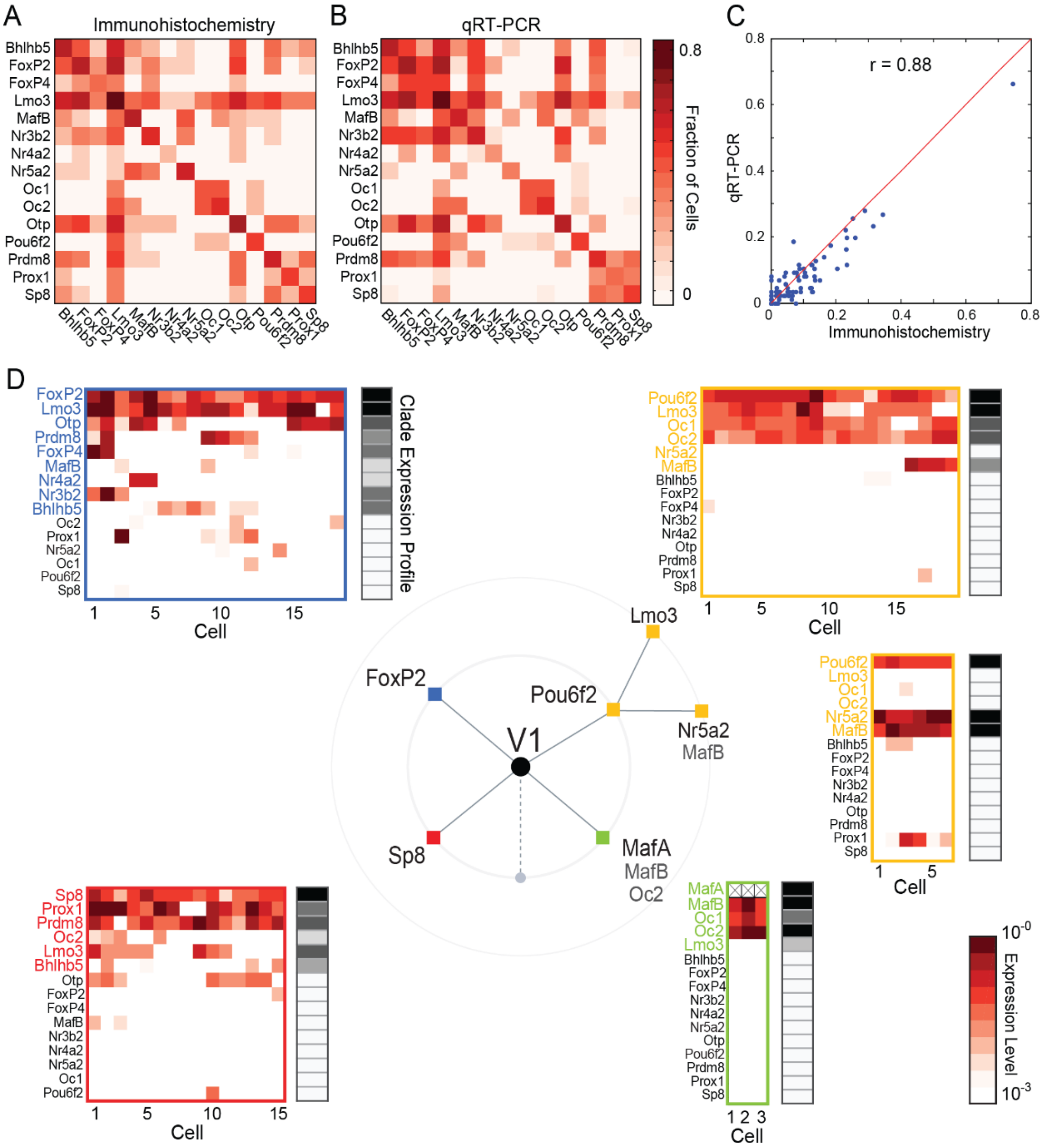

A
Immunohistochemistry
B
qRT-PCR
Bhlhb5
FoxP2
FoxP4
Lmo3
MafB
Nr3b2
Nr4a2
Nr5a2
Oc1
Oc2
Otp
Pou6f2
Prdm8
Prox1
Sp8
0.8
0
Fraction of Cells
C
r = 0.88
qRT-PCR
Immunohistochemistry
0
0.2
0.4
0.6
0.8
D
Clade Expression Profile
Cell
1
5
10
15
V1
FoxP2
Pou6f2
Lmo3
Nr5a2
MafB
MafA
MafB
Oc2
Sp8
MafA
Oc1
Oc2
Lmo3
1 2 3
$10^{-0}$
$10^{-3}$
Expression Level

**Figure 4.1 (preceding page). Single Cell RT-PCR Confirms Antibody Measurements and Validates Candidate Clade Expression Profiles**

(a) Matrix of V1 interneurons representing fraction of cells expressing single and paired TFs. Fractional values of single TFs represented as diagonal elements.

(b) Matrix calculated using RT-PCR cells after relative expression thresholding values (plotted as in (a)); n = 86 cells.

(c) Immunohistochemistry vs qRT-PCR values ((a) vs (b)) show a correlation value of 0.88.

(d) Single cell expression profiles can be arranged according to cladistic analysis; second tier predictions are corroborated in the Pou6f2 clade. The clade expression profile (C.E.P.) was computed by averaging expression profiles of inferred cell types belonging to each clade, weighted by their posterior inclusion probability. These profiles, computed solely from immunohistochemistry data, match the clustered qRT-PCR measurements. 25 of 86 total cells span the remainder (i.e., were not assigned to the clusters shown here; data not shown), consistent with the ratio of remainder cell types.

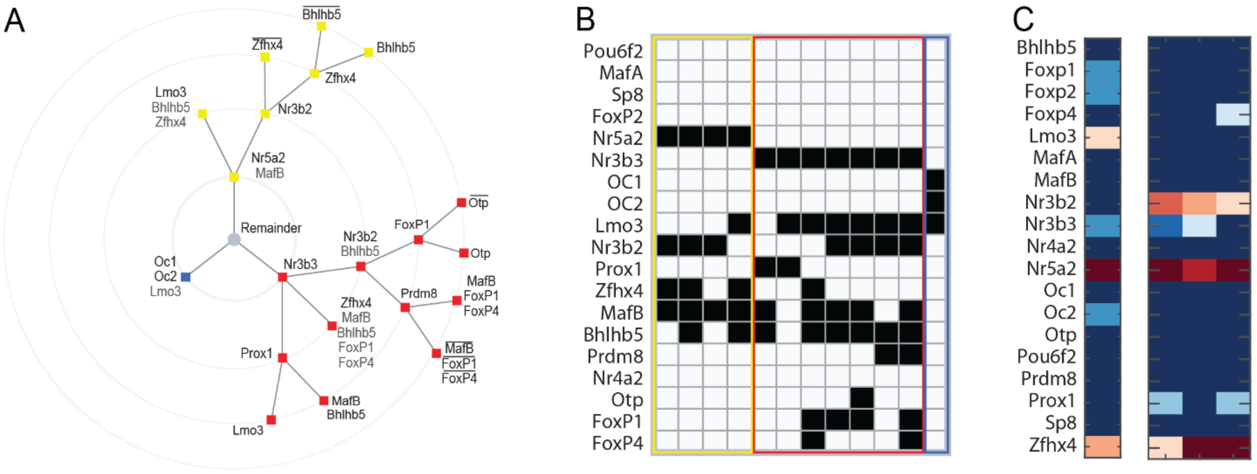


**Figure 4.2 Single Cell RT-PCR and Expression Profiles Lacking Expression of Cladal Genes.**

(a) Clade diagram of cell types lacking expression of any cladal gene. These cell types complete the description set forth in Figure 3.10.

(b) Cell types of cell types lacking expression of any cladal gene. Reproduced as in figure 3.10 for completeness.

(c) Single cells lacking expression of any cladal gene show patterns of expression that might be attribute to the Nr5a2 clade. Because we lack enough coverage, the number of cells belonging to these cell types is too small to formalize any claim. Cells are reproduce here for completeness and they just conform to an anecdotal result.

### *Immunohistochemical Experiments*

Although based solely on single and paired expression data, our results also allow us to make predictions about measurements that involve triple labeling of transcription factors, data not used in the Bayesian analysis. These predictions can be computed by summing the inferred cell type fractions over all cell types expressing the factors being studied. Guided by Bayesian optimal experimental design and antibody availability, we examined expression fractions for the transcription factors Sp8 (Figure 4.3a).The number of possible cell types expressing Sp8 is 18; assessing each of these possibilities individually would be highly laborious (and in general might not be feasible if the required antibodies are not compatible). To explore cell types within this clade more efficiently, we analyzed coincident expression of three transcription factors, Sp8, Prox1 and Prdm8, examining not only populations within the clade but also sampling additional populations outside it (Figure 4.3a-c).

Most of the predicted fractions for the 7 potential combinations of these factors are in good agreement with their measured values (predictions are portrayed in histograms of Figure 4.3b, Table 4.1). The predictions arising from the spatial analysis are more constrained, with smaller standard deviations, and generally more accurate than the non-spatial predictions. Moreover, we validated the predicted absence of the combination Prdm8+, Prox1+, Sp8- (Figure 4.3b). Taken together, these results indicate that the Bayesian approach accurately infers the transcriptional profile of cell types within the parental V1 population.

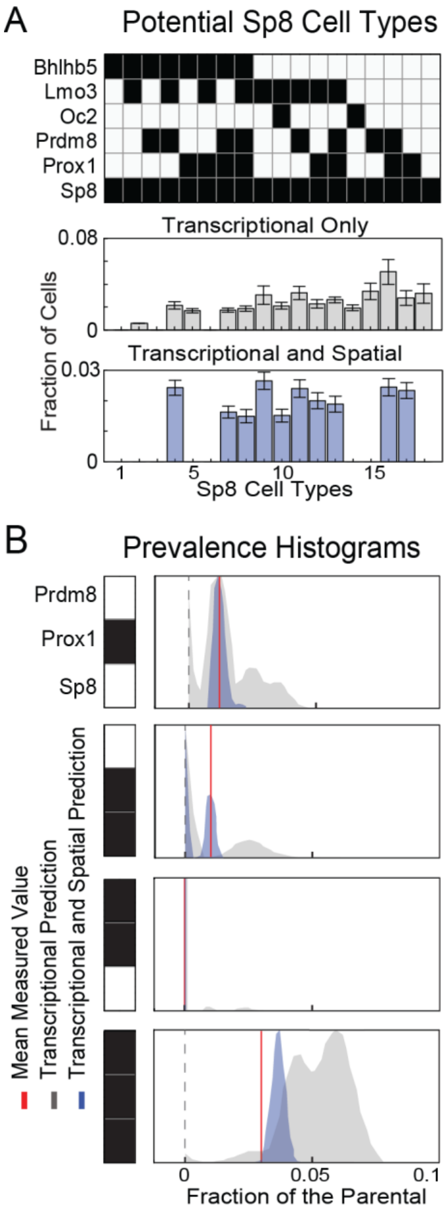


**Figure 4.3. Triple Immunohistochemical measurements validate model predictions.**

(a) 18 Potential cell types expressing Sp8 TF (top), along with their inferred fractions within the parental population (mean ± SD, as in Figure 2A, middle and bottom), calculated using solely TF expression information (middle) or both expression and spatial information (bottom).

(b) Predicted prevalence for measured triplet antibody combinations. Mean measured value is depicted as a red line. Predicted values are indicated in gray or blue (computed using protein expression information only, or with spatial information, respectively).

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Prdm8 | □ | □ | ■ | ■ | □ | ■ | ■ |
| Prox1 | □ | ■ | □ | ■ | ■ | □ | ■ |
| SP8 | ■ | □ | □ | □ | ■ | ■ | ■ |
| Measurement | 0.02±2E-3 | 0.01±3E-3 | 0.13±1E-2 | 1E-3±2E-3 | 0.01±5E-3 | 0.05±8E-3 | 0.03±4E-3 |
| Transcriptional Inference | 0.03±1E-2 | 0.02±1E-2 | 0.11±1E-2 | 3E-3±7E-3 | 0.01±1E-2 | 0.03±1E-2 | 0.05±1E-2 |
| Spatial Transcriptional Inference | 0.01±1E-3 | 0.01±2E-3 | 0.09±5E-3 | 1E-5±8E-5 | 0.01±5E-5 | 0.03±2E-3 | 0.04±2E-3 |

**Table 4.1. Bayesian Inference Accurately Predicts Interneuron Prevalence.**

Measured fraction of expression of Sp8, Prox1 and Prdm8 TFs. For each V1 subpopulation, values indicate measured or predicted prevalence (mean ± SD). To compute the predicted prevalence of each Sp8$^+$ population, predicted fractional values are summed over unobserved TFs Bhlhb5, Oc2 and Lmo3.

Finally, our results also enabled us to test predictions about spatial distributions of neurons expressing pairs of transcription factors, on the basis of spatial information from single transcription factors. Paired transcription factor settling positions are predicted by adding the spatial distributions of inferred cell types expressing the pair under study. Examining the predicted spatial distributions, we focused on cases in which the combination of two transcription factors confined an inferred V1 neuronal type to a highly restricted region of the parental V1 distribution, and for which compatible antibodies were available. We found that our predictions faithfully colocalize with the actual distributions assessed in p0 caudal lumbar spinal segments (Figure 7C-E). In all cases, the centroids of inferred distributions localize within 100 µm of their measured counterparts. These results indicate that the Bayesian approach, by virtue of incorporating dual cellular sources of information, correctly predicts the spatial distribution of novel gene combinations.

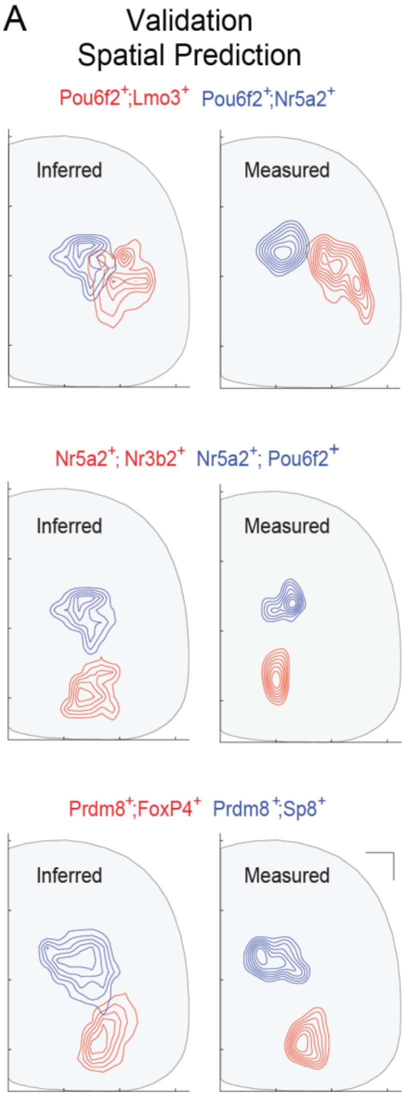


**Figure 4.4. Observed Dual Transcription Factor Spatial Distributions Corroborate Bayesian Estimation.**

Spatial distributions for dual transcription factor-gated V1 subsets can be predicted accurately. Left indicates prediction; right measured distributions.

Scale bar = 100 µm.

*Recapitulating cell type diversity from single cell RNA-seq experiments.*

Lastly, to establish the general applicability of our Bayesian approach, we evaluated its ability to identify cell types in systems where an estimate of cellular diversity had been extracted by other analysis procedures.

We first focused on the zebrafish embryo, for which single cells have been transcriptionally profiled by RNAseq and mapped to their location of origin [Satija et al 2015]. Although the delineation of the entire cellular repertoire was not attempted in that work, the analysis of single cell cluster profiles across the marginal region of the embryo is consistent with seven cell types (Figure 4.5a). We sought to determine whether our sparse Bayesian methods are able to achieve this result given simulated data generated by randomly subsampling the dataset from Satija et al (2015). In the absence of spatial information, the sparse Bayesian algorithm estimated 5 +/- 1 cell types. The transcriptional profile of each inferred cell type corresponds to one of the ground-truth candidates, but our procedure underestimated total cell-type number (Figure 4.5b). Introducing spatial information into the analysis increases the number of correctly inferred cell types to 6 +/- 1 (Figure 4.5c), close to 7, the ground-truth number. Thus as in the V1 study, inference is improved by incorporating additional cellular characteristics – in this case location. The coarse description of the spatial distributions and their similar shapes, together with the random selection of the subset of genes incorporated in our analysis, are likely reasons that the algorithm slightly underestimates cell-type diversity. Nevertheless, the results obtained by the Bayesian approach are generally in good agreement with those obtained by clustering the original RNAseq data.

We next analyzed cortical interneuron diversity, where 16 interneuronal cell types have been identified on the basis of RNAseq data in mouse somatosensory cortex and hippocampal CA1 neurons [Zeisel et al, 2015]. From this data set, single and pairwise measurements were created, with errors assigned to each measurement. We used this dataset to construct Bayesian estimates of neocortical diversity in the absence of spatial information, varying the number of genes used in the analysis, the noise in the measurements and the amount of missing data (representing antibody incompatibility). From this, we inferred 12.7 +/- 0.3 cell types over a range of 13 to 16 genes used, all 12 corresponding to correctly inferred expression profiles (Figure 4.5d). In every example, the sparse Bayesian analysis outperformed the NNCLS approach, which overestimated the number of cell types by nearly 100%. We next used all selected genes to estimate sensitivity to noise and missing data, and observed a larger effect for noise (Figure 4.5e). As the noise level and amount of missing data tend to zero, we correctly infer the total number of cell types and their expression profiles (Figure 4.5f). These analyses establish the sparse Bayesian approach as an effective means of estimating neuronal cell type diversity, and provide further insight into the benefits of incorporating spatial information when obtaining accurate estimates.

Recapitulating, once our statistical framework has been validated we can confidently assert that our Bayesian analysis of V1 molecular diversity and settling position identifies ~50 cell types within this interneuron population localizing in spatially discrete domains. Nevertheless, it remains unclear whether these cell types represent functionally distinct populations that displayed specific and differential electrophysiological properties.

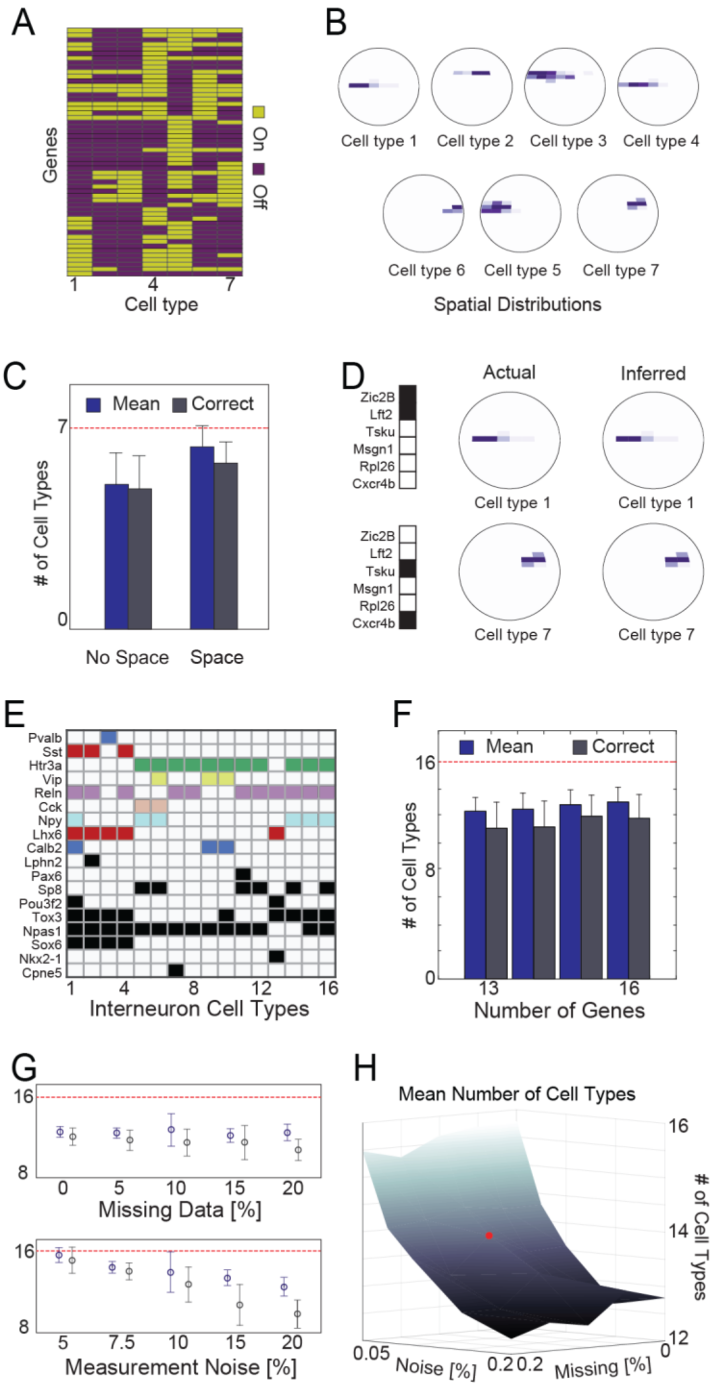


**Figure 4.5. Recapitulating single cell diversity by means of sparse Bayesian regression.**

(a) Expression profiles of zebrafish cell types identified by Satija et al (2015).

(b) Spatial distribution of each cell type in (a).

(c) In blue, posterior mean ± SD number of cell types per sample. In gray, mean ± SD number of correctly identified cell types.

(d) Examples of two correctly inferred spatial distributions.

(e) Interneuronal cell types identified by Zeisel et al (2015). Commonly used markers are color coded as in Zeisel et al (2015) and TF are colored in black.

(f) Sparse Bayesian regression underestimates total cell type. We randomly selected 13 to 16 genes from the list of markers defined in (e). In blue, posterior mean ± SD number of cell types per sample. In gray, mean ± SD number of correctly identified cell types. The expression profile of the first 16 patterns are compared to the true patterns. Red dashed line indicates ground truth value of 16.

(g) Impact of missing data and error in the measurement dataset. (Top) Fixing the measurement error at ten percent and using all the genes described in (e), the performance of the algorithm remains constant when varying the amount of data removed. (Bottom) Fixing the amount of missing data at ten percent and using all the genes described in (e), the performance of the algorithm decreases as the measurement noise approaches 20 percent.

(h) Landscape representing the mean number of inferred cell types when varying the amount of missing data and the noise in the measurements. Red Dot indicates a level of missing data and noise similar to V1 interneurons.

### 4.2.2 Physiological distinctions among V1 clades.

To determine whether molecular distinctions in V1 identity reflect differences in interneuron physiology, we analyzed the electrophysiological properties of neurons in V1$^{FoxP2}$ and V1$^{Pou6f2}$ clades, as well as Renshaw interneurons within V1 (V1$^{R}$), in a spinal cord slice preparation at p10-p14. To label neurons in the V1$^{FoxP2}$ clade, FoxP2::Flpo transgenic mice was used, and used an intersectional genetic strategy in which *En1::Cre; FoxP2::Flpo; RCE.dual.GFP* mice selectively express GFP in V1$^{FoxP2}$ interneurons [Bikoff et al, 2016]. To identify both the V1$^{Pou6f2}$ clade and V1$^{R}$ interneurons, we used *MafB::GFP; En1::Cre; Rosa.lsl.tdT* mice, in which two distinct GFP+/tdT+ V1 subsets could be distinguished: a dorsal subset fully contained within the V1$^{Pou6f2}$ clade, and a ventral subset corresponding to V1$^{R}$ interneurons. Approximately half of all V1$^{R}$ interneurons express MafA, and they serve as a proxy for the V1$^{MafA}$ clade.

We found that V1$^{FoxP2}$, V1$^{Pou6f2}$, and V1$^{R}$ subsets could be distinguished by their passive and active membrane properties (Figure 4.6). At hyperpolarized (< -80 mV) membrane potentials, distinctive active properties included: (i) the prominence of spike after-hyperpolarization (AHP) and early transient low-threshold depolarizations, (ii) the extent of initial spike bursting, and (iii) the degree of spike-frequency adaptation (SFA) during steady-state firing.

Analysis of V1$^{FoxP2}$ interneurons using whole-cell current-clamp recording revealed action potentials with a large and fast-rising AHP, no transient low-threshold depolarizations, no initial spike bursting, and little or no SFA (Figure 4.6a-c). V1$^{Pou6f2}$ interneurons segregated into a lateral bursting subset with a large low-threshold depolarization (Figure 4.6d, e) and a medial non-bursting subset with a much smaller transient depolarization (Figure 4.6g, h). At a molecular level these physiological distinctions likely correspond to the mediolateral positional segregation of

V1$^{Pou6f2/Nr5a2}$ and V1$^{Pou6f2/Lmo3}$ interneurons (Figures 3.10a). Both V1$^{Pou6f2}$ subsets exhibited SFA, likely resulting from the buildup of long duration AHPs during successive spikes (Figure 4.6f, i). V1$^{R}$ interneurons exhibited a large low-threshold depolarization and short AHPs, resulting in a strong bursting phenotype with no evident SFA during steady state firing (Figure 4.6j, l). Thus, V1$^{FoxP2}$, V1$^{Pou6f2}$, and V1$^{R}$ interneurons can be distinguished by their biophysical properties, consistent with their molecular and positional segregation into distinct V1 clades.

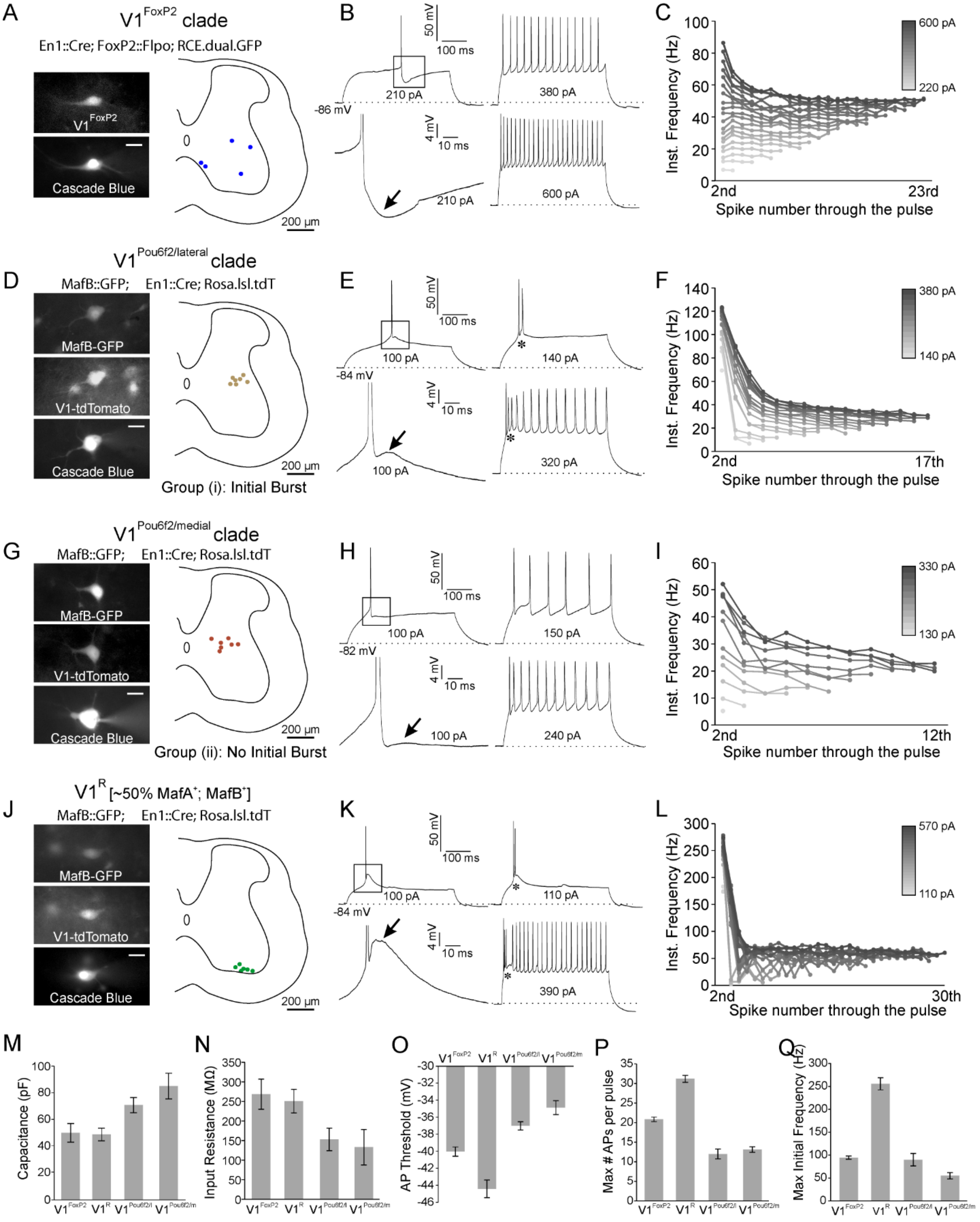

A
$V1^{FoxP2}$ clade
En1::Cre; FoxP2::Flpo; RCE.dual.GFP
$V1^{FoxP2}$
Cascade Blue
0
200 µm
B
50 mV
100 ms
210 pA
380 pA
-86 mV
4 mV
10 ms
210 pA
600 pA
C
Inst. Frequency (Hz)
100
80
60
40
20
0
2nd
23rd
Spike number through the pulse
600 pA
220 pA
$V1^{Pou6f2/lateral}$ clade
D
MafB::GFP; En1::Cre; Rosa.lsl.tdT
MafB-GFP
V1-tdTomato
Cascade Blue
0
200 µm
Group (i): Initial Burst
E
50 mV
100 ms
100 pA
140 pA
-84 mV
4 mV
10 ms
100 pA
320 pA
F
Inst. Frequency (Hz)
140
120
100
80
60
40
20
0
2nd
17th
Spike number through the pulse
380 pA
140 pA
$V1^{Pou6f2/medial}$ clade
G
MafB::GFP; En1::Cre; Rosa.lsl.tdT
MafB-GFP
V1-tdTomato
Cascade Blue
0
200 µm
Group (ii): No Initial Burst
H
50 mV
100 ms
100 pA
150 pA
-82 mV
4 mV
10 ms
100 pA
240 pA
I
Inst. Frequency (Hz)
60
50
40
30
20
10
0
2nd
12th
Spike number through the pulse
330 pA
130 pA
$V1^{R}$ [~50% $MafA^{+}$; $MafB^{+}$]
J
MafB::GFP; En1::Cre; Rosa.lsl.tdT
MafB-GFP
V1-tdTomato
Cascade Blue
0
200 µm
K
50 mV
100 ms
100 pA
110 pA
-84 mV
4 mV
10 ms
390 pA
L
Inst. Frequency (Hz)
300
250
200
150
100
50
0
2nd
30th
Spike number through the pulse
570 pA
110 pA
M
Capacitance (pF)
100
80
60
40
20
0
$V1^{FoxP2}$ $V1^{R}$ $V1^{Pou6f2/l}$ $V1^{Pou6f2/m}$
N
Input Resistance (MΩ)
350
300
250
200
150
100
50
0
$V1^{FoxP2}$ $V1^{R}$ $V1^{Pou6f2/l}$ $V1^{Pou6f2/m}$
O
AP Threshold (mV)
-30
-32
-34
-36
-38
-40
-42
-44
-46
$V1^{FoxP2}$ $V1^{R}$ $V1^{Pou6f2/l}$ $V1^{Pou6f2/m}$
P
Max # APs per pulse
35
30
25
20
15
10
5
0
$V1^{FoxP2}$ $V1^{R}$ $V1^{Pou6f2/l}$ $V1^{Pou6f2/m}$
Q
Max Initial Frequency (Hz)
300
250
200
150
100
50
0
$V1^{FoxP2}$ $V1^{R}$ $V1^{Pou6f2/l}$ $V1^{Pou6f2/m}$

**Figure 4.6 (preceding page). Electrophysiological Characterization of V1 Clades**
(a- c) Physiology of V1$^{FoxP2}$ interneurons. (a) V1$^{FoxP2}$ interneurons (n = 5) targeted for recording and filled with Cascade Blue in En1::Cre; FoxP2::Flpo; RCE.dual.GFP mice.
Scale bar = 20 µm.
(b) Firing properties of V1$^{FoxP2}$ cells show a prominent after-hyperpolarization (AHP, arrow), a non-bursting phenotype, and an absence of spike frequency adaptation (SFA).
(c) Instantaneous firing (IF) frequency for each action potential (dot) through pulses of increasing current amplitudes (20 pA steps). Little or no SFA is observed below 460 pA.

(d- f) Physiology of V1$^{Pou6f2/lateral}$ interneurons. (d) Position of V1$^{Pou6f2/lateral}$ interneurons (n = 7) in MafB::GFP; En1::Cre; Rosa.lsl.tdT mice.
(e) Transient low-threshold depolarization (arrow), with an initial burst (asterisks), and the presence of SFA throughout the pulse.
(f) SFA, indicated by the decreasing instantaneous frequency of successive action potentials.

(g- i) Physiology of V1$^{Pou6f2/medial}$ interneurons. (g) Position of V1$^{Pou6f2/medial}$ interneurons (n = 7).
(h) Neurons exhibit a non-burst phenotype and a weak low-threshold depolarization (arrow, H).
(i) IF plot showing SFA.

(j-k) Physiology of V1R interneurons, representing the V1MafA clade. (j) Position of V1R interneurons (n = 6) in En1::Cre; Rosa.lsl.tdT; MafB::GFP mice.
(k) Neurons show prominent low-threshold depolarization (arrow), and burst firing (asterisks).
(l) IF plot shows absence of SFA.

(m-p) Passive electrophysiological properties
(m-n) and action potential threshold (G) for V1$^{FoxP2}$ (n = 7), V1$^{Pou6f2/lateral}$ (n = 7), V1$^{Pou6f2/medial}$ (n = 7), and V1$^{R}$ (n = 6) interneurons. The V1$^{Pou6f2}$ subset exhibited significantly larger membrane capacitance (Cm), and trended towards a lower input resistance (Ri) (although this did not reach significance). V1$^{Pou6f2}$ interneurons also showed a significantly more depolarized action potential threshold, when compared to V1$^{Foxp2}$ and V1$^{R}$ interneurons. Cm, $p < 0.001$, one-way ANOVA; Bonferroni post-hoc test: $p < 0.02$, V1$^{Pou6f2/medial}$ vs V1$^{R}$ or V1$^{FoxP2}$; AP Threshold, $p < 0.001$, one-way ANOVA; Bonferroni post-hoc test: $p < 0.001$, V1$^{Pou6f2/lateral}$ vs V1$^{R}$, V1$^{Pou6f2/medial}$ vs V1$^{R}$, V1$^{Pou6f2/medial}$ vs V1$^{FoxP2}$; $p < 0.05$, V1$^{Pou6f2/lateral}$ vs V1$^{FoxP2}$. This suggests that V1$^{Pou6f2}$ interneurons likely require larger summation of incoming synaptic inputs to reach firing threshold than V1Foxp2 and V1R interneurons.

(q) Maximum number of action potentials per 435 msec pulse; $p < 0.001$, one-way ANOVA; Bonferroni post-hoc test: $p < 0.001$ for all pairwise comparisons except V1$^{Pou6f2/lateral}$ vs V1$^{Pou6f2/medial}$ ($p > 0.05$). (I) V1R interneurons displayed the highest initial firing frequency. $p < 0.001$, one-way ANOVA; Bonferroni post-hoc test: $p < 0.001$, V1$^{R}$ vs V1$^{Pou6f2/lateral}$, V1$^{Pou6f2/medial}$, and V1$^{FoxP2}$.

## 4.3 Discussion

*Validation of Sparse Bayesian Regression*

Many previous studies have proposed computational methods to reveal cellular diversity. Some of these methods do not represent the general biological instance and cannot be applied when there is no knowledge of the underlying cell type expression profiles. A different set of methods require a large number of cells to estimate the true extent of cellular diversity. The former methodologies utilize a genome wide approach, which assess the entire cellular transcriptome at single cell resolution. These methods suffer from the recognized disconnection between mRNA and protein expression patterns [Gygi *et al*, 1999]; [Vogel & Marcotte, 2012], highlighting the necessity of analysis at the protein expression level. Furthermore, no method calculates neither the uncertainty of estimates, nor the integration of different sources of information -two characteristics that have proven critical to ultimately discover underlying cell heterogeneity.

The experiments presented in this chapter support our Bayesian framework as an effective method to reveal cellular diversity. Quantitative PCR provide a general validation, demonstrating the overall organization of predicted diversity. However, only two cell types were fully validated with this method, indicating that many more cells would have been needed to fully delineate the entire transcriptional catalog. A recent study suggests the necessity of profiling more than five thousands cell types to fully outline V1 heterogeneity (Tasic et al, 2016).

Triple immunohistochemical experiments were used to validate the ability of our method to infer the existence and prevalence of cell types. Although many cell types are not delineated by just a few antibodies, the selected combinations allowed us to probe some individual cell types even corroborating the predicted absence of a population. In addition, spatial measurements of pairs of transcription factors provide an estimate of the algorithm's capability to correctly estimate the spatial location of cell predictions. Further experiments need to be performed to substantiate predictions of more complicated combinations, such as the ones at the edges of the clade diagram. In addition, computational experiments inspired on biological data validated the ability of our Bayesian methodology to infer expression profiles. The results from these experiments demonstrate that while predictions based solely on transcriptional information sometimes underestimate total diversity, computationally predicted expression profiles faithfully reproduce underlying diversity.

### *Characterizing diversity with orthogonal cellular properties*

In previous chapters we have defined V1 cell types based on their molecular profile and their location in space, validating predictions according to these cellular features. Nonetheless, this definition lacked independent evidence that the identified V1 subsets represent distinct functional cell types. In fact, if transcriptionally distinct V1 subsets indeed represent different cell types, a reasonable expectation is that they show different electrophysiological characteristics.

In this chapter we describe the result of a collaboration with Dr. Francisco Alvarez at Emory University, in which we explored whether different V1 subsets are associated with distinctive electrophysiological characteristics. Answering this necessitated the use of several transgenic mouse lines to facilitate the identification of different V1 subsets for targeted recordings. Electrophysiological analysis of three V1 subsets - V1$^{FoxP2}$, V1$^{Pou6f2}$, and V1$^{R}$ interneurons - demonstrated clear functional differences amongst them, consistent with their molecular and positional segregation into distinct V1 clades. We characterized the presence or absence of a bursting phenotype, spike frequency adaptation (SFA), a low-threshold depolarization underlying the spike, the nature of the after-hyperpolarization (AHP), and passive electrophysiological properties including Ri (input resistance) and Cm (cell capacitance).

Electrophysiological characterization resulted in three major findings: (i) we find that the $V1^{FoxP2}$ subset is characterized by a non-bursting phenotype, the absence of significant SFA, a rapid and large AHP, and no low-threshold depolarization. (ii) $V1^{R}$ interneurons exhibit a bursting phenotype, the absence of SFA, a slower AHP, and a large low-threshold depolarization. (iii) $V1^{Pou6f2}$ interneurons differ from both $V1^{FoxP2}$ and $V1^{R}$ interneurons in exhibiting SFA, and in addition they can be fractionated into bursting and non-bursting subsets with lateral and medial positional biases.

The addition of these new data provide an independent functional parameter that further supports the notion emerging from our molecular, positional, and circuit-level analysis - that V1 subsets reflect meaningful distinctions in cell identity and function.

# Chapter 5

# Spinal Inhibitory Interneuronal Micro Circuitry

## 5.1 Introduction

In previous chapters we show that transcriptionally defined V1 interneuron subsets possess distinct settling position and electrophysiological properties. This chapter represents combined effort on experimental procedures performed by Jay Bikoff in which we study what is the relationship of spinal inhibitory interneuronal circuits and motor pools. Motivated by the observed stereotypical spatial segregation exhibited by interneuronal populations we wondered, is this cellular feature used to direct the establishment of interneuronal circuits? To answer this question we focus on two distinct populations, Renshaw and Sp8 expressing interneurons and study their input and output connections in association with motor pools innervated by three different sets of muscles.

By fluorescently labelling the different subsets and performing cholera toxin tracing experiments, we show that different interneuron micro circuitry is established for motor pools innervating hip, ankle, and foot muscles - revealing joint-selective variation in circuitry. The absence of a single, canonical circuit wiring diagram suggests that spinal inhibitory circuits are individually tailored to meet the biomechanical demands of particular joints and limb muscles.

## 5.2 Results

### 5.2.1 Mapping the relative position of V1 subpopulations and motor pools

The stereotypic nature of the V1 positional matrix led us to examine whether the spinal motor system takes advantage of spatial segregation in the construction of inhibitory microcircuits. To address the functional implications of V1 positional diversity we focused on the local connectivity of V1 subpopulations that settle at dorsal and ventral extremes of the parental V1 domain. Ventrally, we examined Renshaw (V1$^{R}$) interneurons, a subtype that mediates recurrent inhibition of motor neurons [Eccles et al., 1961]. V1$^{R}$ interneurons express MafB, Oc1/2, and the calcium binding protein calbindin, and are included, in part, within the MafA clade [Carr et al., 1998]; [Stam et al., 2012]. As a dorsal comparator population we examined neurons in the V1$^{Sp8}$ clade, which consists of 8 inferred V1 cell types (Figure 2.7).

To mark V1$^{Sp8}$ interneurons, we employed an intersectional and inducible genetic strategy, using a Sp8::flpoERT2 transgenic mouse line that evades the early expression of Sp8 in neuronal progenitors [Bikoff et al, 2016]. The use of Sp8::flpoERT2 and En1::cre driver lines crossed with an RCE.dual.GFP reporter allele, combined with tamoxifen administration at p0, resulted in selective labeling of a cluster of V1$^{Sp8}$ interneurons in the dorsomedial region of the parental V1 domain (Figure 5.1A). To mark V1$^{R}$ interneurons we took advantage of their V1 derivation and expression of calbindin, permitting their identification in En1::cre; Rosa.lsl.eGFP mice. We examined p21 mice, an age at which spinal and descending circuits have reached sufficient maturity to direct adult-like patterns of locomotion (Clarke and Still, 2001). This analysis revealed a compact V1$^{R}$ settling domain, just medial to lateral motor column (LMC) motor neurons, and close to the ventral border of the gray matter (Figure 5.1b). Thus, these genetic labeling strategies define molecularly and positionally distinct V1 subpopulations.

We focused on three hindlimb motor pools to probe the organizational features of inhibitory microcircuits that control limb musculature. Gluteus (GL) motor neurons innervate hip extensor muscles and occupy an extreme ventral position in the LMC. Tibialis anterior (TA) and instrinsic foot (IF) motor neurons innervate ankle flexor and foot plantar-flexor muscles respectively, and occupy a similar dorsal position within the LMC [McHanwell and Biscoe, 1981]. Assessment of $V1^{R}$ and $V1^{Sp8}$ positions with respect to GL, TA, and IF motor pools, retrogradely labeled by cholera toxin B (CTB) muscle injection in p21 En1::Cre; Rosa.lsl.FP (tdT or eGFP) mice, indicated that $V1^{R}$ interneurons occupy a ventral position close to that of GL motor neurons, whereas $V1^{Sp8}$ interneurons occupy a dorsal position close to that of TA and IF motor neurons (Figure 5.1c-e). The dorsal and ventral positioning of these two V1 populations provided a spatial reference point for assessing two elements of spinal motor microcircuitry: (i) the nature of motor input to V1 interneuron sets and (ii) the organization of V1 interneuron output onto discrete motor neuron pools.

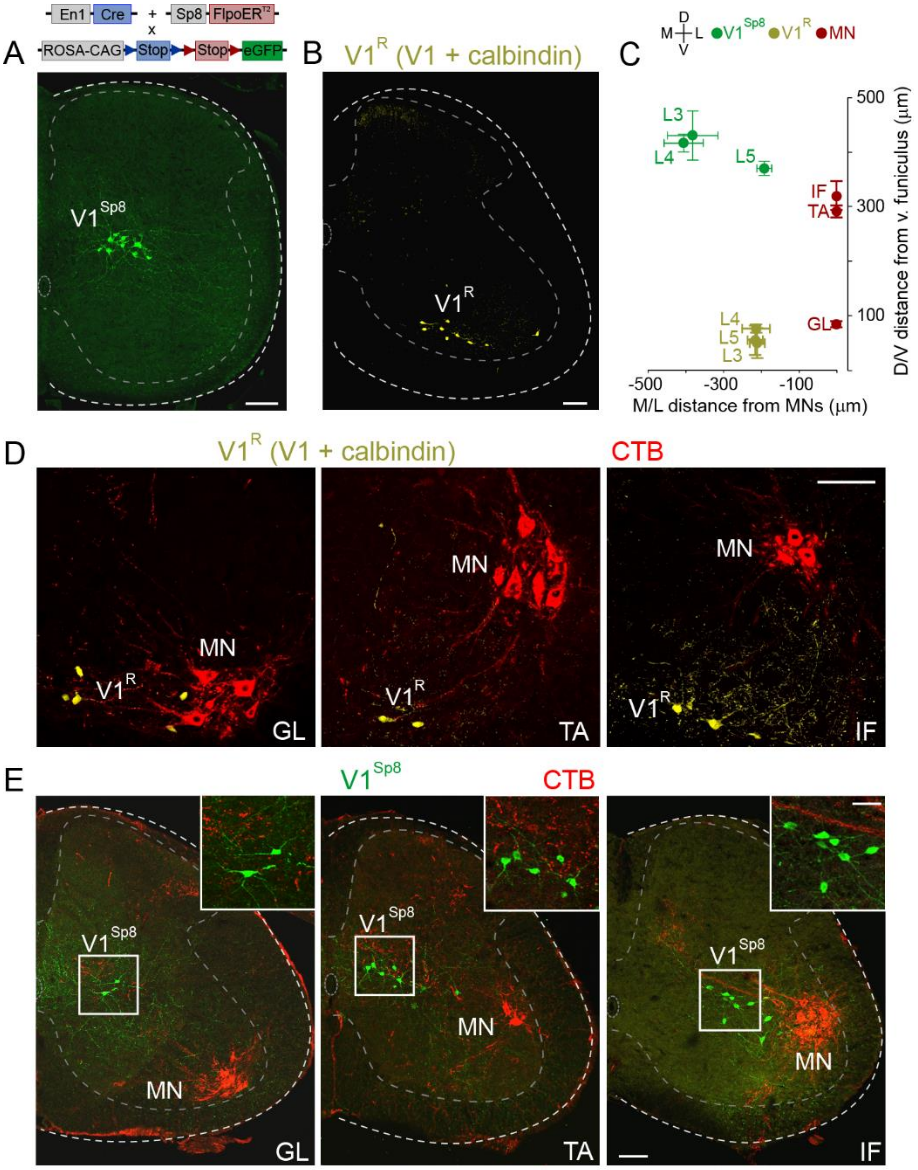

En1 Cre + Sp8 FlpoER^T2
x
ROSA-CAG Stop Stop eGFP
A
V1^Sp8
B
V1^R (V1 + calbindin)
V1^R
C
D
M L
V
V1^Sp8
V1^R
MN
L3
L4
L5
IF
TA
GL
L4
L5
L3
500
300
100
-500
-300
-100
D/V distance from v. funiculus (µm)
M/L distance from MNs (µm)
D
V1^R (V1 + calbindin)
CTB
MN
V1^R
GL
MN
V1^R
TA
MN
V1^R
IF
E
V1^Sp8
CTB
V1^Sp8
MN
GL
V1^Sp8
MN
TA
V1^Sp8
MN
IF

**Figure 5.1 (preceding page). Relative Position of V1R and V1Sp8 Interneurons to Motor Pools**

(a) V1Sp8 interneurons (green), in p12 lumbar spinal cord of En1::Cre; Sp8::FlpoERT2; RCE.dual.GFP mice.

(B) V1R interneurons (yellow, colocalization mask of eGFP and calbindin immunoreactivity) in ~p21 En1::Cre; RCE.lsl.GFP lumbar spinal cord.

(c) V1R and V1Sp8 position with respect to GL, TA, and IF motor pools in ~p21 mice. Motor pool D/V positions: GL: 84 ± 3 µm, TA: 291 ± 6 µm, IF: 321 ± 15 µm, from dorsal border of ventral funiculus.

(d) D/V position of V1R interneurons (yellow) with respect to CTB-backfilled GL, TA, and IF motor pools (MN, red) in ~p21 lumbar spinal cord. D/V distances: V1R ventral to GL, TA, and IF motor neurons by 8 ± 3 µm, 242 ± 14 µm, and 264 ± 13 µm, respectively. $p < 0.0001$, oneway ANOVA; Bonferroni post-hoc test: $p < 0.001$, TA or IF vs GL. M/L distances were not significantly different ($p = 0.99$, one-way ANOVA).

(e) D/V position of V1Sp8 interneurons (green) with respect to CTB-backfilled GL, TA, and IF motor pools (red) in ~p21 lumbar spinal cord. V1Sp8 dorsal to GL, TA, and IF by 332 ± 8 µm, 139 ± 23 µm, and 50 ± 8 µm, respectively ($p < 0.0001$, one-way ANOVA; Bonferroni post-hoc test: $p < 0.001$, TA or IF vs GL; $p < 0.05$, TA vs IF). In the M/L axis, V1Sp8 interneurons were significantly closer to IF than to GL or TA (192 ± 11 µm versus 406 ± 26 µm or 382 ± 33 µm, respectively; $p < 0.01$, one-way ANOVA; Bonferroni post-hoc test; $p < 0.01$, IF vs GL or TA). Values are mean ± SEM, $n \geq 3$ animals per condition. Scale bars = 100 µm or 50 µm (inset).

### 5.2.2 Positional constraints on interconnectivity between V1 interneurons and motor neurons

We next considered whether the settling position of V1$^{R}$ and V1$^{Sp8}$ interneurons predicts their interconnectivity with motor neurons, analyzing first the extent to which V1$^{R}$ and V1$^{Sp8}$ interneurons receive motor neuron collateral input. Analysis of the location of CTB+; vAChT+ motor axon collateral terminals indicated that GL and TA, but not IF motor axons form dense collateral arbors that are confined to an extreme ventral domain, overlapping the position of V1$^{R}$ interneurons, and ventral to V1$^{Sp8}$ interneurons. By implication, V1$^{R}$ but not V1$^{Sp8}$ interneurons are positioned to receive synaptic input from GL and TA but not IF motor neurons.

To assess motor neuron collateral input to V1$^{R}$ or V1$^{Sp8}$ interneurons, CTB was injected into individual muscles and analyzed in p21 En1::cre; RCE.lsl.GFP or En1::cre; Sp8::flpoERT2; RCE.dual.GFP tamoxifen-treated mice. The density of CTB-labeled, vAChT+ boutons was determined on the soma and proximal dendrites of V1$^{R}$ (Figure 5.2b, c) or V1$^{Sp8}$ interneurons (Figure 5.2d, e). GL and TA motor neurons both provided synaptic input to V1$^{R}$ interneurons, with each innervating about half of all V1$^{R}$ interneurons at comparable CTB+; vAChT+ bouton densities (Figure 5.2c). Nevertheless, V1$^{R}$ interneurons did not receive collateral input from IF motor neurons (Figure 5.2c). Moreover, motor neuron collateral innervation was restricted to V1$^{R}$ interneurons, with V1$^{Sp8}$ interneurons receiving no motor axon collateral contacts (Figure 5.2d, e), consistent with their dorsal position. Thus, the different ventral and dorsal positions of V1$^{R}$ and V1$^{Sp8}$ interneurons predict the status of motor as well as sensory neuron input.

These findings prompted us to examine whether subsets of V1 interneurons provide differential input to motor pools, in a manner that conforms to their settling position. We analyzed the connections of V1$^{R}$ or V1$^{Sp8}$ interneurons with GL, TA, and IF motor neurons (Figure 5.2f). V1-derived, calbindin+ terminals from V1$^{R}$ interneurons contacted GL and TA motor neurons at similar densities in ~p21 En1::cre; Rosa.lsl.tdT or En1::cre; RCE.lsl.GFP mice (Figure 5.2g, h; p = 0.94 or 0.84 for dendrites and soma, respectively, by two-tailed Student's t-test). Thus, ventrally located V1$^{R}$ interneurons target motor pools with similar efficacy at different dorsoventral locations. In contrast, IF motor neurons exhibited little input from V1$^{R}$ neurons (Figure 5.2g, h), despite the similar dorsoventral settling position of TA and IF motor pools. The incidence of V1$^{R}$ innervation was ~7-fold greater on TA proximal dendrites and ~14-fold greater on TA soma, by comparison with IF motor neurons (Figure 5.2h). The ventral position of V1R interneurons therefore does not appear to constrain their ability to innervate motor neurons at different dorsoventral positions.

We also explored whether the dorsal settling position of V1$^{Sp8}$ interneurons limits their connectivity with motor neurons. In En1::cre; Sp8::flpoERT2; RCE.dual.GFP mice, V1$^{Sp8}$ interneurons provided sparse and uniform contacts with GL, TA, and IF motor neurons (Figure 5.2i, j). The innervation density of V1$^{Sp8}$ inputs was <20% that of V1R inputs onto GL or TA motor neurons - although this value is likely to underestimate the actual incidence of motor neuron innervation, because only ~30% of V1$^{Sp8}$ interneurons are labeled after tamoxifen exposure. These observations indicate that the dorsomedial positioning of V1$^{Sp8}$ interneurons does not limit their ability to innervate diverse motor neuron targets. Together, these data suggest that the settling position of inhibitory interneurons is a determinant of input, but not output, connectivity in spinal motor microcircuits.

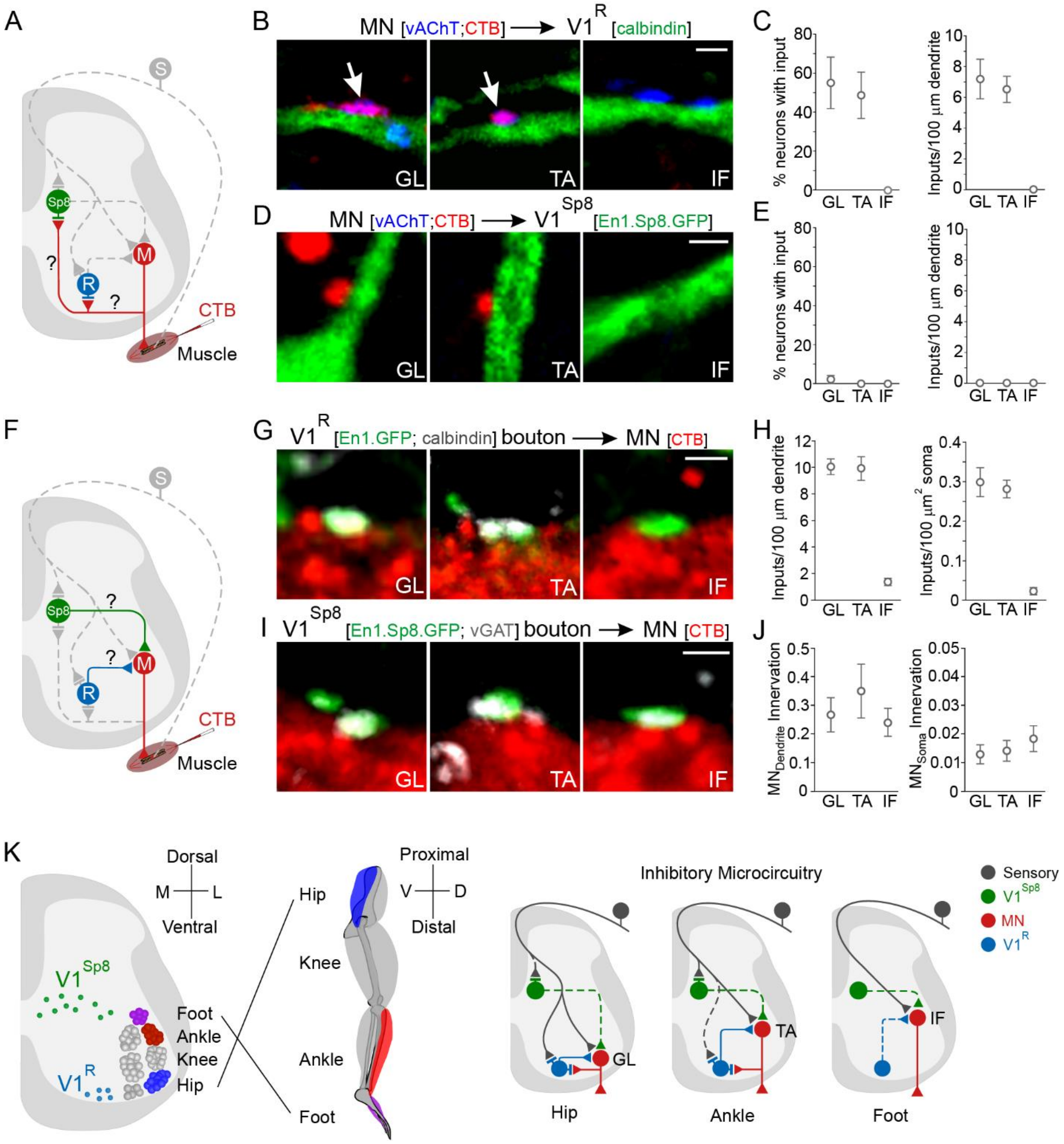

A
S
Sp8
?
R
?
M
CTB
Muscle
B
MN [vAChT;CTB] ⟶ V1 R [calbindin]
GL
TA
IF
C
% neurons with input
80
60
40
20
0
GL TA IF
Inputs/100 μm dendrite
10
8
6
4
2
0
GL TA IF
D
MN [vAChT;CTB] ⟶ V1 Sp8 [En1.Sp8.GFP]
GL
TA
IF
E
% neurons with input
80
60
40
20
0
GL TA IF
Inputs/100 μm dendrite
10
8
6
4
2
0
GL TA IF
F
S
Sp8
?
?
M
R
CTB
Muscle
G
V1 R [En1.GFP; calbindin] bouton ⟶ MN [CTB]
GL
TA
IF
H
Inputs/100 μm dendrite
12
10
8
6
4
2
0
GL TA IF
Inputs/100 μm² soma
0.4
0.3
0.2
0.1
0
GL TA IF
I
V1 Sp8 [En1.Sp8.GFP; vGAT] bouton ⟶ MN [CTB]
GL
TA
IF
J
MNDendrite Innervation
0.5
0.4
0.3
0.2
0.1
0
GL TA IF
MNSoma Innervation
0.05
0.04
0.03
0.02
0.01
0
GL TA IF
K
Dorsal
M
L
Ventral
V1 Sp8
V1 R
Foot
Ankle
Knee
Hip
Proximal
V
D
Distal
Hip
Knee
Ankle
Foot
Inhibitory Microcircuitry
GL
Hip
TA
Ankle
IF
Foot
● Sensory
● V1 Sp8
● MN
● V1 R

**Figure 5.2 (preceding page). Specificity of Interneuron-Motor Neuron Interconnectivity at Individual Joints**

(a) Assay of pool-specific motor input to interneurons.

(b- c) V1R interneurons receive CTB+; vAChT+ input from GL and TA (arrows) but not IF motor neurons (MN). Scale bar = 2 µm.

(c) Left, V1R interneurons with input from GL, TA, or IF MNs. $p = 0.02$, one-way ANOVA; Bonferroni post-hoc test: $p < 0.05$, GL or TA vs IF. Right, CTB+ MN inputs/100 µm of V1R dendrite length. $p = 0.002$, one-way ANOVA; Bonferroni post-hoc test: $p < 0.01$, GL or TA vs IF; $p > 0.5$, GL vs TA, $n \geq 3$ animals, and 23 (GL), 24 (TA), or 15 (IF) cells.

(d- e) Absence of MN input to V1Sp8 interneurons. GL, $n = 4$ animals, 43 cells; TA, $n = 2$ animals, 52 cells; IF, $n = 3$ animals, 43 cells.

(f) Assay of interneuron input onto motor pools.

(g- h) V1R interneurons preferentially innervate GL and TA relative to IF motor pools, on proximal MN dendrites (h, left) or soma (h, right). $p < 0.0001$, one-way ANOVA; Bonferroni post hoc test: $p < 0.001$, GL or TA vs IF, $n = 4$ animals, and 31 (GL), 21 (TA), or 27 (IF) cells.

(i- j) V1Sp8 interneurons sparsely and uniformly innervate motor pools acting on different joints. Number of V1Sp8 inputs/100 µm MN dendrite or 100 µm2 of soma area, normalized to V1Sp8 interneuron number. $p = 0.53$ or 0.65 for dendrites and soma, respectively, one-way ANOVA, $n \geq 3$ animals, 35 (GL), 42 (TA), or 59 (IF) cells. Scale bars = 2 µm. All data are mean ± SEM. (K) V1R and V1Sp8 microcircuits operating on hip, ankle, and foot motor neurons. Solid and dotted lines represent prevalent and sparse synaptic connectivity

# 5.3 Discussion

*Position as a Determinant in the Organization of Inhibitory Microcircuits*

The relevance of neuronal settling position in spinal connectivity has emerged from studies on the synaptic organization of sensory connections with motor neurons. Proprioceptive afferents target distinct dorsoventral domains of the ventral spinal cord in a manner independent of motor neuron character [Sürmeli et al., 2011], and thus the stereotypy of settling position is needed for the formation of selective sensory connections. Similarly, $V1^R$ interneurons receive input from ventrally projecting hip afferents, whereas dorsal $V1^{Sp8}$ interneurons receive input both from dorsally-directed ankle afferents as well as from hip afferents. Thus, V1 positional stereotypy has implications for motor microcircuit organization in the realm of input selectivity.

The finding that $V1^R$ interneurons receive selective input from hip muscle afferents sheds light on a long-standing uncertainty about the status of sensory input to $V1^R$ interneurons. Classical studies in cat focused on sensory feedback from knee and ankle muscles, and argued for the absence of functional monosynaptic sensory connectivity with $V1^R$ interneurons (Ryall and Piercey, 1971). Later studies in rodent spinal cord, however, provided physiological evidence for direct sensory input to $V1^R$ interneurons during early postnatal development [Mentis et al., 2006]. These divergent conclusions can be reconciled through an appreciation of the dominance of proprioceptive input from hip afferents, an afferent source not examined in cat. Nevertheless, the extent to which this circuit functions at later developmental stages is unclear because the strength of sensory inputs to $V1^R$ interneurons decreases in the adult [Mentis et al., 2006]

The density of hip afferent inputs to $V1^R$ interneurons presumably forms a disynaptic feedforward inhibitory pathway to motor neurons, in addition to the role of $V1^R$ interneurons in recurrent inhibition. Sensory-evoked feedforward inhibition could modulate the temporal features and dynamic range of excitatory responses of hip motor neurons, as with inhibitory interneurons in hippocampal and cortical circuits [Pouille et al., 2009]. An inhibitory signal dependent on hip position could also modulate flexion/extension transitions during the step cycle [McVea et al., 2005] and/or reflex actions at the ankle joint [Knikou and Rymer, 2002].

The link between interneuron settling position and microcircuit wiring is so far largely correlative. Nevertheless our data, combined with previous findings on the relevance of motor neuron positioning [Sürmeli et al., 2011], supports the view that the precision of interneuron location constrains circuit wiring. The role of neuronal settling position in organizing interneuron circuits appears restricted to input connectivity. V1 interneuron position is not predictive of motor pool target connections, reminiscent of observations that motor neuron settling position is not required for the innervation of specific limb muscles [Demireva et al., 2011].

Positional constraints are likely to act in conjunction with molecular recognition systems in defining final connectivity profiles. Precedent for such recognition systems has emerged from analysis of repellant sema3e-plexinD1 signaling in sensory-motor connectivity [Fukuhara et al., 2013]; [Pecho-Vrieseling et al., 2009]. The existence of repellent cues could explain how the dorsal termination zone of IF sensory afferents is not associated with direct synaptic contact with $V1^{Sp8}$ interneurons, despite the proximity of presynaptic axons and $V1^{Sp8}$ dendrites. In addition, the extent of interneuron dendritic arborization could relieve constraints on input connectivity imposed by somatic clustering. Although the dendritic arbors of $V1^{R}$ interneurons are largely confined to the ventral spinal cord [Lagerback and Kellerth, 1985], $V1^{Sp8}$ interneurons exhibit larger dendritic arbors (J. Bikoff, unpublished observation), potentially expanding synaptic input. Thus it seems unlikely that position alone directs input connectivity with V1 subsets.

# 6 Conclusion

In this thesis, we developed statistical tools to explore the organization of cellular populations and use them to discover unparalleled diversity within an interneuronal population. Through a focus on V1 interneurons, we identified nineteen transcription factors whose combinatorial expression helped us describe extensive diversity within this inhibitory set. Different V1 subsets have distinct physiological characteristics and occupy stereotypic clustered positions in the ventral spinal cord. This genetically defined spatial plan has predictive relevance for inhibitory microcircuit organization. Indeed, variant V1 microcircuits are used to control motor pools that innervate muscles at different limb joints, documenting the absence of a fixed circuit architecture for interneurons that control limb movement.

### *Interneuron Diversity and its Implications for Motor Control*

V1 interneurons comprise a highly diverse set of transcriptionally distinct neuronal types, posing questions about the purpose of such heterogeneity. Diversity may reflect the demand that interneurons receive varied inputs from numerous sources. The activity of motor neurons is regulated by over a dozen supraspinal neuronal systems (Lemon, 2008), many of which engage only a restricted set of all possible motor pools: thus rubrospinal input is restricted to motor pools controlling distal muscles, and vestibulospinal input to motor pools innervating extensor muscles [Grillner and Hongo, 1972]; [McCurdy et al., 1987]. These descending systems presumably engage interneurons with a selectivity that matches the specificity of motor neuron recruitment.

Distinct subsets of V1 interneurons may therefore be recruited by different descending systems so as to link sensory input with intermediary descending control pathways. The high degree of V1 transcriptional diversity could provide a means of establishing distinctions in settling position or molecular recognition cues that facilitate the integration of multiple input systems and output modules.

The heterogeneity exhibited by V1 interneurons is likely to extend to other spinal interneuron populations. Small subsets of spinal V0 interneurons have been delineated on the basis of selective profiles of transcription factor expression, best exemplified by a compact cluster of Pitx2+ V0c interneurons that represent the source of cholinergic C-bouton inputs to motor neurons [Zagoraiou et al., 2009]. Moreover, many of the transcription factors that delineate V1 subsets are expressed by small subsets of inhibitory V2b and excitatory V2a interneurons, raising the possibility that conserved elements of input and output wiring specificity are encoded by a common set of transcription factors within different excitatory and inhibitory interneuron sets. If the extent of diversity of V1 interneurons extends to each cardinal (V0, V2a/b, and V3) interneuron population [Francius et al., 2013], the fidelity of motor output could depend on the coordinated activity of > 200 subsets of ventral interneurons.

It remains unclear whether the diversity evident in V1 interneurons has predictive relevance for other CNS circuits. The spinal motor system could require a greater degree of interneuron diversification than the brain, because of the last-order and non-redundant nature of motor neuron output and the behavioral imperative to confer precise patterns of muscle activation. Nevertheless, the predictive view may be nearer the mark. Single-cell transcriptional profiling from interneurons in primary somatosensory cortex and CA1 hippocampus have revealed at least sixteen different subsets, with the potential for yet greater diversity [Zeisel et al., 2015]. In addition, many of the transcription factors that delineate subsets of V1 interneurons are expressed by subsets of cortical interneurons [Tasic et al., 2016]. Thus, it is likely that principles of spinal interneuron heterogeneity and function have relevance for circuit organization and function in the brain.

### *Broader Implications of a Bayesian Analysis of Cellular Diversity*

Our statistical approach has relevance well beyond a focus on spinal V1 interneurons, and could prove useful in further delineating neuronal cell types elsewhere in the nervous system. Cortical projection neurons fractionate into a few broad classes based on patterns of target innervation and distinctions in gene expression, yet the extent to which any single broad class of pyramidal neurons is itself heterogeneous remains unclear [Greig et al, 2013]. The classification of interneuron cell types in the brain has proven particularly challenging [Ascoli, 2008]; [DeFelipe et al, 2013]; [Kepecs and Fishell, 2014], although studies of hippocampal interneuron diversity suggest a degree of heterogeneity that approaches that found for spinal V1 interneurons. Within CA1 hippocampus, over 20 inhibitory interneuron subtypes have been identified, based on anatomical, molecular, or electrophysiological distinctions [Krook-Magnusen et al, 2012]. Singlecell transcriptome analysis of primary somatosensory cortex or CA1 hippocampus interneurons has identified 16 molecularly distinct interneuron cell types, which likely represents a lower bound on diversity [Zeisel et al, 2015]. Thus, insight into interneuronal diversity in the spinal cord may inform studies to address heterogeneity throughout the brain. Our analysis also has implications for genetic strategies aimed at manipulating circuit elements throughout the nervous system. The minimal number of TFs needed to define a single V1 cell type uniquely has been identified on the basis of clade profiles and is, on average, 4 ± 1. This indicates that individual

TFs are generally not sufficient to isolate V1 neuronal types, consistent with findings in other neuronal systems [Sanes and Masland, 2015]. The difficulty in identifying single TFs that uniquely define a cell type may reflect the prevalence of combinatorial TF codes [Philippidou and Dasen, 2015], and could explain the difficulty in delineating individual motor neuron pools [De Marco Garcia and Jessell, 2008].

*Future directions*

The variant circuit architectures exhibited by V1 interneurons may be a general feature of spinal motor microcircuits. The anatomical and physiological characterization of the circuitry of $V1^R$ interneurons is consistent with physiological descriptions in cat of a reduced degree of recurrent inhibition for motor neurons that innervate distal compared to proximal limb musculature [Illert and Wietelmann, 1989]; [McCurdy and Hamm, 1992]. Thus, recurrent inhibition is not implemented uniformly across motor pools. Local motor microcircuits are therefore differentially tailored to the workings of individual muscles. We are pursuing a continuation of this research extending circuit characterization to thoracic and cervical populations addressing: how are thoracic and cervical populations organized? And are there specific markers defining subsets of thoracic and cervical cell types?

The field of transcriptome profiling has been revolutionized in the last five years by the emergence of detailed protocols for library preparations, lowering cost of deep sequencing techniques and new algorithmic tools that allow the assessment of gene expression with enough sensitivity to perform it at the single cell level. However, due to a reliance on heuristic or inappropriate dimensionality reduction techniques, no robust statistical technique to analyze these data and discover underlying cell types has yet emerged. As a consequence, the assignment of cells to their corresponding cell types is done without any measure of statistical confidence. Our Bayesian statistical framework can be extended taking advantage of recent advances in machine learning and stochastic processes to model single-cell data. Such methods will permit a completely unsupervised modelling of the clustering process while allowing enough flexibility to represent the data collection process and its nuances in detail. The ideas developed in this thesis should allow genome wide methods to incorporate different cellular characteristics, quantify uncertainty of estimates, and make predictions to guide further experiments.

# Bibliography

Abbas A.R., Wolslegel K., Seshasayee D., Modrusan Z., Clark H.F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS ONE 7, e6098.

Alaynick W.A., Jessell T.M., Pfaff S.L. (2011). SnapShot: spinal cord development. Cell 146, 178.

Alvarez F.J., Jonas P.C., Sapir T., Hartley R., Berrocal M.C., Geiman E.J., Todd A.J., Goulding M. (2005). Postnatal phenotype and localization of spinal cord V1 derived interneurons. J. Comp. Neurol. 493, 177-192.

Alvarez F.J., Fyffe R.E. (2007). The continuing case for the Renshaw cell. J. Physiol. 584, 31-45.

Amamoto R., Arlotta P. (2014). Development-inspired reprogramming of the mammalian central nervous system. Science 343, 1239882.

Arber, S. (2012). Motor circuits in action: specification, connectivity, and function. Neuron 74, 975-989.

Armañanzas R., Ascoli G.A. (2015). Towards the automatic classification of neurons. Trends Neurosci. 5, 307-318.

Ascoli G.A. (2008). Neuroinformatics grand challenges. Neuroinformatics. I, 1-3.

Augen J. (2005). Bioinformatics in the Post-Genomic Era: Genome, Transcriptome, Proteome, and Information-Based Medicine. (Addison-Wesley Press).

Bazot C., Dobigeon N., Tourneret J.Y., Zaas A.K., Ginsburg G.S., Hero A.O. (2013). Unsupervised Bayesian linear unmixing of gene expression microarrays. BMC Bioinformatics 14:99.

Benito-Gonzalez, A., and Alvarez, F.J. (2012). Renshaw cells and Ia inhibitory interneurons are generated at different times from p1 progenitors and differentiate shortly after exiting the cell cycle. J. Neurosci. 32, 1156-1170.

Betley, J.N., Wright, C.V., Kawaguchi, Y., Erdelyi, F., Szabo, G., Jessell, T.M., and Kaltschmidt, J.A. (2009). Stringent specificity in the construction of a GABAergic presynaptic inhibitory circuit. Cell 139, 161-174.

Bikoff J.B., Gabitto M.I., Rivard A.F., Drobac E., Machado T.A., Miri A., Brenner-Morton S., Famojure E., Diaz C., Alvarez F.J., et al. (2016). Spinal inhibitory interneuron diversity delineates variant motor microcircuits. Cell.

Blacklaws J, Deska-Gauthier D, Jones CT, Petracca YL, Liu M, Zhang H, Fawcett JP, Glover JC, Lanuza GM, Zhang Y. (2015). Sim1 is required for the migration and axonal projections of V3 interneurons in the developing mouse spinal cord. Dev. Neurobiol. 75, 1003-1017.

Borowska J., Jones C.T., Zhang H., Blacklaws J., Goulding M., Zhang Y. (2013). Functional subpopulations of V3 interneurons in the mature mouse spinal cord. J Neurosci. 33, 18553-18565.

Briscoe, J., Pierani, A., Jessell, T.M., and Ericson, J. (2000). A homeodomain protein code specifies progenitor cell identity and neuronal fate in the ventral neural tube. Cell 101, 435-445.

Brown, A.G. (1981). Organization in the Spinal Cord: The Anatomy and Physiology of Identified Neurones (New York: Springer-Verlag).

Brownstone R.M., Bui T.V. (2010). Spinal interneurons providing input to the final common path during locomotion. Prog. Brain Res. 187, 81-95.

Bühlmann P., Rütimann P, Van de Geer S, Zhang C. (2013). Correlated variables in regression: Clustering and sparse estimation. Journal of Statistical Planning and Inference 143, 1835–1858.

Carr, P.A., Alvarez, F.J., Leman, E.A., and Fyffe, R.E. (1998). Calbindin D28k expression in immunohistochemically identified Renshaw cells. Neuroreport 9, 2657-2661.

Cond´e, F., Lund, J. S., Jacobowitz, D. M., Baimbridge, K. G. and Lewis, D. A. (1994). Local circuit neurons immunoreactive for calretinin, calbindin DâAR28k or parvalbumin in monkey prefronatal cortex: Distribution and morphology. The Journal of comparative neurology 341, 95–116.

Crone S.A., Quinlan K.A., Zagoraiou L., Droho S., Restrepo C.E., Lundfald L., Endo T., Setlak J., Jessell T.M., Kiehn O., Sharma K. (2008). Genetic ablation of V2a ipsilateral interneurons disrupts left-right locomotor coordination in mammalian spinal cord. Neuron 60, 70-83.

Dalla Torre di Sanguinetto S.A., Dasen J.S., Arber S. (2008) Transcriptional mechanisms controlling motor neuron diversity and connectivity. Curr. Opin. Neurobiol. 1, 36-43.

Dasen J.S., Tice B.C., Brenner-Morton S., Jessell T.M. (2005) A Hox regulatory network establishes motor neuron pool identity and target-muscle connectivity. Cell 3, 477-91.

Dasen J.S., Jessell T.M. (2009). Hox networks and the origins of motor neuron diversity. Curr. Top. Dev. Biol. 88, 169-200.

DeFelipe J., López-Cruz P.L., Benavides-Piccione R., Bielza C., Larrañaga P., Anderson S., Burkhalter A., Cauli B., Fairén A., Feldmeyer D., et al (2013). New insights into the classification and nomenclature of cortical GABAergic interneurons. Nat. Rev. Neurosci. 14, 202-216.

DeFelipe J., Hendry S.H., Jones E.G. (1989). Visualization of chandelier cell axons by parvalbumin immunoreactivity in monkey cerebral cortex. Proc. Natl. Acad. Sci. USA, 86, 2093-2097.

DeFelipe J., Hendry S.H., Jones E.G. (1989). Synapses of double bouquet cells in monkey cerebral cortex visualized by calbindin immunoreactivity. Brain Res. 503, 49-54.

DeFelipe, J., Hendry, S. H., Hashikawa, T., Molinari, M. and Jones, E. G. (1990). A microcolumnar structure of monkey cerebral cortex revealed by immunocytochemical studies of double bouquet cell axons. NSC 37, 655–673.

DeFelipe J., López-Cruz P.L., Benavides-Piccione R., Bielza C., Larrañaga P., Anderson S., Burkhalter A., Cauli B., Fairén A., Feldmeyer D., et al. (2013). New insights into the classification and nomenclature of cortical GABAergic interneurons. Nat. Rev. Neurosci. 14, 202-216.

del R´ıo, M. R. and DeFelipe, J. (1995). A light and electron microscopic study of calbindin D-28k immunoreactive double bouquet cells in the human temporal cortex. Brain research 690, 133–140.

Del Barrio M.G., Taveira-Marques R., Muroyama Y., Yuk D.I., Li S., Wines-Samuelson M., Shen J., Smith H.K., Xiang M., Rowitch D., Richardson W.D. (2007). A regulatory network involving Foxn4, Mash1 and delta-like 4/Notch1 generates V2a and V2b spinal interneurons from a common progenitor pool. Development 134, 3427-3436.

Dehorter N., Ciceri G., Bartolini G., Lim L., Pino I, Marín O. (2015). Tuning of fast-spiking interneuron properties by an activity-dependent transcriptional switch. Science 349, 1216-1220.

De Marco Garcia N.V., Jessell T.M. (2008). Early motor neuron pool identity and muscle nerve trajectory defined by postmitotic restrictions in Nkx6.1 activity. Neuron 57, 217-231.

Demireva, E.Y., Shapiro, L.S., Jessell, T.M., and Zampieri, N. (2011). Motor neuron position and topographic order imposed by β- and γ-catenin activities. Cell 147, 641-652.

Eccles, J.C., Eccles, R.M., Iggo, A., and Ito, M. (1961). Distribution of recurrent inhibition among motoneurones. J. Physiol. 159, 479-499.

Erkkilä T., Lehmusvaara S., Ruusuvuori P., Visakorpi T., Shmulevich I., Lähdesmäki H. (2010). Probabilistic analysis of gene expression measurements from heterogeneous tissues. Bioinformatics 20, 2571-2587.

Fishell G., Rudy B. (2011). Mechanisms of inhibition within the telencephalon: "where the wild things are". Annu. Rev. Neurosci. 34, 535-567.

Francius, C., Harris, A., Rucchin, V., Hendricks, T.J., Stam, F.J., Barber, M., Kurek, D., Grosveld, F.G., Pierani, A., Goulding, M., et al. (2013). Identification of multiple subsets of ventral interneurons and differential distribution along the rostrocaudal axis of the developing spinal cord. PLoS ONE 8, e70325.

Fonseca M., Soriano E., Ferrer I., Martinez A., Tuñon T. (1993). Chandelier cell axons identified by parvalbumin-immunoreactivity in the normal human temporal cortex and in Alzheimer's disease. Neuroscience 55, 1107-1116.

Fukuhara, K., Imai, F., Ladle, D.R., Katayama, K., Leslie, J.R., Arber, S., Jessell, T.M., and Yoshida, Y. (2013). Specificity of monosynaptic sensory-motor connections imposed by repellent Sema3E-PlexinD1 signaling. Cell Rep. 5, 748-758.

Fyffe R.E. (1990). Evidence for separate morphological classes of Renshaw cells in the cat's spinal cord. Brain Res. 1, 301-304.

Gabitto, M.I., Pakman, A., Bikoff, J.B., Abbott, L.F., Jessell, T.M., and Paninski, L. (2016). Bayesian sparse regression analysis documents the diversity of spinal inhibitory interneurons. Cell.

Gabbott, P. L. and Bacon, S. J. (1996). Local circuit neurons in the medial prefrontal cortex (areas 24a,b,c, 25 and 32) in the monkey: I. Cell morphology and morphometrics. The Journal of comparative neurology 364, 567–608.Kubota et al., 1994;

Gaujoux R., Seoighe C. (2012). Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. Infect. Genet. Evol. 5, 913-921.

Gelman A., Carlin J.B., Stern H. S., Dunson D.B., Vehtari A. (2013). Bayesian Data Analysis. Boca Raton, FL: Chapman & Hall/CRC.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. J. Amer. Stat. Assoc. 88, 881–889.

Glasgow, S.M., Henke, R.M., Macdonald, R.J., Wright, C.V., and Johnson, J.E. (2005). Ptf1a determines GABAergic over glutamatergic neuronal cell fate in the spinal cord dorsal horn. Development 132, 5461-5469.

Gong T., Hartmann N., Kohane I.S., Brinkmann V., Staedtler F., Letzkus M., Bongiovanni S., Szustakowski J.D. (2011). Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PLoS ONE 11, e27156.

Gonzalez-Forero, D., and Alvarez, F.J. (2005). Differential postnatal maturation of GABAA, glycine receptor, and mixed synaptic currents in Renshaw cells and ventral spinal interneurons. J. Neurosci. 25, 2010-2023.

Gosgnach, S., Lanuza, G.M., Butt, S.J., Saueressig, H., Zhang, Y., Velasquez, T., Riethmacher, D., Callaway, E.M., Kiehn, O., and Goulding, M. (2006). V1 spinal neurons regulate the speed of vertebrate locomotor outputs. Nature 440, 215-219.

Goulding M, Lanuza G, Sapir T, Narayan S. (2002). The formation of sensorimotor circuits. Curr. Opin. Neurobiol. 12, 508-515.

Goulding M. (2009). Circuits controlling vertebrate locomotion: moving in a new direction. Nat. Rev. Neurosci. 10, 507-518.

Goulding, M., Bourane, S., Garcia-Campmany, L., Dalet, A., and Koch, S. (2014). Inhibition downunder: an update from the spinal cord. Curr. Opin. Neurobiol. 26, 161-166.

Grange P., Bohland J., Okaty B., Sugino K., Bokil H., Nelson S., Ng L., Hawrylycz M., Mitra P. (2014) Cell-type-based model explaining coexpression patterns of genes in the brain. Proc. Nat. Acad. Sci. 111, 5397-5402.

Greig L.C., Woodworth M.B., Galazo M.J., Padmanabhan H., Macklis J.D. (2013). Molecular logic of neocortical projection neuron specification, development and diversity. Nat. Rev. Neurosci. 11, 755-769.

Grillner, S., and Hongo, T. (1972). Vestibulospinal effects on motoneurones and interneurones in the lumbosacral cord. Prog. Brain Res. 37, 243-262.

Grillner S., Kozlov A., Kotaleski J.H. (2005). Integrative neuroscience: linking levels of analyses. Curr. Opin. Neurobiol. 15, 614-621.

Grün D, van Oudenaarden A. (2015). Design and Analysis of Single-Cell Sequencing Experiments. Cell 163, 799-810.

Gygi S.P., Rochon Y., Franza B.R., Aebersold R. (1999). Correlation between protein and mRNA abundance in yeast. Mol. Cell Biol. 3, 1720-1730.

Hendry S.H., Jones E.G., Emson P.C., Lawson D.E., Heizmann C.W., Streit P. (1989). Two classes of cortical GABA neurons defined by differential calcium binding protein immunoreactivities. Exp. Brain. Res. 76, 467-472.

Hughes DI, Mackie M, Nagy GG, Riddell JS, Maxwell DJ, Szabó G, Erdélyi F, Veress G, Szucs P, Antal M, Todd AJ. (2005). P boutons in lamina IX of the rodent spinal cord express high levels of glutamic acid decarboxylase-65 and originate from cells in deep medial dorsal horn. Proc. Natl. Acad. Sci. USA 102, 9038-9043.

Hultborn H., Jankowska E., Lindström. (1971). Recurrent inhibition of interneurones monosynaptically activated from group Ia afferents. J Physiol. 3, 613-636.

Isaacson J.S., Scanziani M. (2011). How inhibition shapes cortical activity. Neuron 2, 231-243.

Insel TR, Landis SC, Collins FS. (2013). Research priorities. The NIH BRAIN Initiative. Science 340, 687-688.

Ishwaran and J.S. Rao. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. Ann. Statist., 33, 730–773.

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

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

Jankowska E. (2001). Spinal interneuronal systems: identification, multifunctional character and reconfigurations in mammals. J. Physiol. 533, 31-40.

Joshi K., Lee S., Lee B., Lee J.W., Lee S.K. (2009). LMO4 controls the balance between excitatory and inhibitory spinal V2 interneurons. Neuron 61, 839-851.

Kandel E., Schwartz J., Jessell T.M., Siegelbaum S. (2012). McGraw-Hill Professional Publishing.

Kawaguchi, Y. and Kubota, Y. (1996). Physiological and morphological identification of somatostatin- or vasoactive intestinal polypeptide-containing cells among GABAergic cell subtypes in rat frontal cortex. The Journal of neuroscience 16, 2701–2715.

Kepecs A., Fishell G. (2014). Interneuron cell types are fit to function. Nature 505, 318-326.

Kjaerulff O., Kiehn O. (1996). Distribution of networks generating and coordinating locomotor activity in the neonatal rat spinal cord in vitro: a lesion study. J. Neurosci. 18, 5777-5794.

Kiehn O. (2011). Development and functional organization of spinal locomotor circuits. Curr. Opin. Neurobiol. 21, 100-109.

Knikou, M., and Rymer, Z. (2002). Effects of changes in hip joint angle on H-reflex excitability in humans. Exp. Brain Res. 143, 149-159.

Kohwi M, Doe CQ. (2013). Temporal fate specification and neural progenitor competence during development. Nat. Rev. Neurosci. 12, 823-838.

Krook-Magnuson E., Varga C., Lee S.H., Soltesz I. (2012). New dimensions of interneuronal specialization unmasked by principal cell heterogeneity. Trends Neurosci. 35, 175-184.

Lagerback, P.A., and Kellerth, J.O. (1985). Light microscopic observations on cat Renshaw cells after intracellular staining with horseradish peroxidase. II. The cell bodies and dendrites. J. Comp. Neurol. 240, 368-376.

Lallemend F., Ernfors P. (2012). Molecular interactions underlying the specification of sensory neurons. Trends Neurosci. 35, 373-381.

Lanuza GM, Gosgnach S, Pierani A, Jessell TM, Goulding M. (2004). Genetic identification of spinal interneurons that coordinate left-right locomotor activity necessary for walking movements. Neuron 42, 375-386.

Lee S., Lee B., Joshi K., Pfaff S.L., Lee J.W., Lee S.K. (2008). A regulatory network to segregate the identity of neuronal subtypes. Dev. Cell. 14, 877-889.

Lemon, R.N. (2008). Descending pathways in motor control. Annu. Rev. Neurosci. 31, 195218.

Liebner D.A., Huang K., Parvin J.D. (2014). MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. Bioinformatics 5, 682-689.

Lin Y., Bloodgood B.L., Hauser J.L., Lapan A.D., Koon A.C., Kim T.K., Hu L.S., Malik A.N., Greenberg M.E. (2008). Activity-dependent regulation of inhibitory synapse development by Npas4. Nature 455, 1198 - 1204.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 5, 1202-1214.

Masland R.H. (2004). Neuronal cell types. Curr. Biol. 14, R497-500.

McCurdy, M.L., Hansma, D.I., Houk, J.C., and Gibson, A.R. (1987). Selective projections from the cat red nucleus to digit motor neurons. J. Comp. Neurol. 265, 367-379.

McCurdy, M.L., and Hamm, T.M. (1992). Recurrent collaterals of motoneurons projecting to distal muscles in the cat hindlimb. J. Neurophysiol. 67, 1359-1366.

McHanwell, S., and Biscoe, T.J. (1981). The localization of motoneurons supplying the hindlimb muscles of the mouse. Philos. Trans. R. Soc. of Lond. B, Biol. Sci. 293, 477-508.

McVea, D.A., Donelan, J.M., Tachibana, A., and Pearson, K.G. (2005). A role for hip position in initiating the swing-to-stance transition in walking cats. J. Neurophysiol. 94, 3497-3508.

Mentis, G.Z., Siembab, V.C., Zerda, R., O'Donovan, M.J., and Alvarez, F.J. (2006). Primary afferent synapses on developing and adult Renshaw cells. J. Neurosci. 26, 13297-13310.

Miri A, Azim E, Jessell TM. (2013). Edging toward entelechy in motor control. Neuron 80, 827-834.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. J. Am. Stat. Assoc. 83, 1023–1032.

Molyneaux B.J., Arlotta P., Macklis J.D. (2007). Molecular development of corticospinal motor neuron circuitry. Novartis Found. Symp. 288, 3-15.

Moran-Rivard L, Kagawa T, Saueressig H, Gross MK, Burrill J, Goulding M. (2001). Evx1 is a postmitotic determinant of v0 interneuron identity in the spinal cord. Neuron 29, 385-399.

Okigawa S., Mizoguchi T., Okano M., Tanaka H., Isoda M., Jiang Y.J., Suster M., Higashijima S., Kawakami K., Itoh M. (2014). Different combinations of Notch ligands and receptors regulate V2 interneuron progenitor proliferation and V2a/V2b cell fate determination. Dev. Biol. 391, 196-206.

Orlovsky G.N., Deliagina T.G., S. Grillner S. (1999). Neuronal Control of Locomotion: From Mollusc to Man. Oxford Neuroscience.

Pakman A. and Paninski L. (2013). Auxiliary-variable exact Hamiltonian Monte Carlo samplers for binary distributions. Advances in Neural Information Processing Systems, 2490-2498.

Pakman A. and Paninski L. (2014). Exact Hamiltonian Monte Carlo for truncated Multivariate Gaussian. Journal of Computational and Graphical Statistics. 23, 2.

Panayi H., Panayiotou E., Orford M., Genethliou N., Mean R., Lapathitis G., Li S., Xiang M., Kessaris N., Richardson W.D., Malas S. (2010). Sox1 is required for the specification of a novel p2-derived interneuron subtype in the mouse ventral spinal cord. J. Neurosci. 30, 12274-12280.

Pecho-Vrieseling, E., Sigrist, M., Yoshida, Y., Jessell, T.M., and Arber, S. (2009). Specificity of sensory-motor connections encoded by Sema3e-Plxnd1 recognition. Nature 459, 842-846.

Philippidou P., Dasen J.S. (2015). Sensory-Motor Circuits: Hox Genes Get in Touch. Neuron 88, 437-440.

Pierani A., Moran-Rivard L., Sunshine M.J., Littman D.R., Goulding M., Jessell T.M. (2001). Control of interneuron fate in the developing spinal cord by the progenitor homeodomain protein Dbx1. Neuron 29, 367-384.

Pouille, F., Marin-Burgin, A., Adesnik, H., Atallah, B.V., and Scanziani, M. (2009). Input normalization by global feedforward inhibition expands cortical dynamic range. Nat. Neurosci. 12, 1577-1585.

Repsilber D., Kern S., Telaar A., Walzl G., Black G.F., Selbig J., Parida S.K., Kaufmann S.H., Jacobsen M. (2010). Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. BMC Bioinformatics 11, 27.

Rossignol S., Dubuc R., Gossard J.P. (2006). Dynamic sensorimotor interactions in locomotion. Physiol. Rev. 86, 89-154.

Ryall, R.W., and Piercey, M.F. (1971). Excitation and inhibition of Renshaw cells by impulses in peripheral afferent nerve fibers. J. Neurophysiol. 34, 242-251.

Sanes J.R., Masland R.H. (2015). The Types of Retinal Ganglion Cells: Current Status and Implications for Neuronal Classification. Annu. Rev. Neurosci. 38, 221-246.

Sapir T., Geiman E.J., Wang Z., Velasquez T., Mitsui S., Yoshihara Y., Frank E., Alvarez F.J., Goulding M. (2004). Pax6 and Engrailed 1 regulate two distinct aspects of Renshaw cell development. J. Neurosci. 5, 1255-1264.

Satija R., Farrell JA., Gennert D., Schier AF, Regev A. (2015) Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol., 33, 495–502.

Saueressig H., Burrill J., Goulding M. (1999). Engrailed-1 and netrin-1 regulate axon pathfinding by association interneurons that project to motor neurons. Development 19, 4201-4212.

Schomburg E.D. (1990). Spinal sensorimotor systems and their supraspinal control. Neurosci Res. 7, 265-340.

Sharma K., Schmitt S., Bergner C.G., Tyanova S., Kannaiyan N., Manrique-Hoyos N., Kongi K., Cantuti L., Hanisch U.K., Philips M.A., et al, (2015). Cell type– and brain region–resolved mouse brain proteome. Nat. Neurosci., 18, 1819 - 1831.

Sharpee T.O. (2014). Toward functional classification of neuronal types. Neuron 83, 1329-1334.

Shen-Orr S., Gaujoux R. (2013) Computational deconvolution: extracting cell type-specific information from heterogeneous samples, Curr. Opin. Immunol. 5, 571-578.

Siegert S., Scherf B.G., Del Punta K., Didkovsky N., Heintz N., Roska B. (2009). Genetic address book for retinal cell types. Nat. Neurosci. 9, 1197-1204.

Smiley, J. F., McGinnis, J. P. and Javitt, D. C. (2000). Nitric oxide synthase interneurons in the monkey cerebral cortex are subsets of the somatostatin, neuropeptide Y, and calbindin cells. Brain research 863, 205–212.

Stam F.J., Hendricks T.J., Zhang J., Geiman E.J., Francius C., Labosky P.A., Clotman F., Goulding M. (2012). Renshaw cell interneuron specialization is controlled by a temporally restricted transcription factor program. Development 131, 179 – 190.

Stepien A.E., Arber S. (2008). Probing the locomotor conundrum: descending the 'V' interneuron ladder. Neuron 60, 1-4.

Stepien A.E., Tripodi M., Arber S. (2010). Monosynaptic rabies virus reveals premotor network organization and synaptic specificity of cholinergic partition cells. Neuron 68, 456-472.

Sürmeli, G., Akay, T., Ippolito, G.C., Tucker, P.W., and Jessell, T.M. (2011). Patterns of spinal sensory-motor connectivity prescribed by a dorsoventral positional template. Cell 147, 653665.

Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., et al. (2016). Adult mouse cortical cell taxonomy revealed by singel cell transcriptomics. Nat. Neurosci. doi:10.1038/nn.4216.

Thomas R.C., Wilson V.J. (1965). Precise localization of Renshaw cells with a new marking technique. Nature 980, 211-213.

Tripodi M., Stepien A.E., Arber S. (2011). Motor antagonism exposed by spatial segregation and timing of neurogenesis. Nature 479, 61-66.

Usoskin D., Furlan A., Islam S., Abdo H., Lönnerberg P., Lou D., Hjerling-Leffler J., Haeggström J., Kharchenko O., Kharchenko P.V., Linnarsson S., Ernfors P. (2015) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat. Neurosci. 18, 145-153.

Vogel C., Marcotte E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. 13, 227-232.

Vrieseling, E., and Arber, S. (2006). Target-induced transcriptional control of dendritic patterning and connectivity in motor neurons by the ETS gene Pea3. Cell 127, 1439-1452.

Wang M., Master S.R., Chodosh L.A. (2006). Computational expression deconvolution in a complex mammalian organ. BMC Bioinformatics. 7, 328.

Windhorst, U. (2007). Muscle proprioceptive feedback and spinal networks. Brain Res. Bull. 73, 155-202.

Zagoraiou, L., Akay, T., Martin, J.F., Brownstone, R.M., Jessell, T.M., and Miles, G.B. (2009). A cluster of cholinergic premotor interneurons modulates mouse locomotor activity. Neuron 64, 645-662.

Zeisel, A., Munoz-Manchado, A.B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138-1142

Zhang J., Lanuza G.M., Britz O., Wang Z., Siembab V.C., Zhang Y., Velasquez T., Alvarez F.J., Frank E., Goulding M. (2014) V1 and V2b interneurons secure the alternating flexor-extensor motor activity mice require for limbed locomotion. Neuron 1, 138-150.

Zhang Y, Narayan S, Geiman E, Lanuza GM, Velasquez T, Shanks B, Akay T, Dyck J, Pearson K, Gosgnach S, Fan CM, Goulding M. (2008). V3 spinal neurons establish a robust and balanced locomotor rhythm during walking. Neuron 60, 84-96.

Zhong Y, Wan YW, Pang K, Chow LM, Liu Z. (2013). Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics 14, 89.

Zuk O., Amir A., Zeisel A., Shamir O., and Shental N. (2013). Accurate profiling of microbial communities from massively parallel sequencing using convex optimization. String Processing and Information Retrieval.

# Studying the neural substrate of nestmate recognition in ants.

## A.1 Introduction

This section is presented as reference for future studies. Ants live in one of the most amazing complex societies from the animal kingdom, achieving one of the highest levels of organization, termed eusociality (greek: eu “truly”, social). Their lives are organized around the maintenance of the colony, sacrificing their individuality in pursuit of the group wellbeing [Holldobler & Willson, 2008]. Depending on the ant species considered, each colony can contain from ten to as many as twenty million members. Interactions among members of these superorganisms are regulated by secreted chemical cues that convey a variety of messages, from the presence of an intruder to the location of a food source. Even more, the recognition of nestmates is also communicated through an array of chemicals which reside on the ants’ cuticles. In this appendix, we investigate the development of techniques aimed to explore the neural substrate mediating the recognition of chemical signals in the ant brain.

### *Behavior*

A hallmark among ant societies is the ability of individuals to recognize members of the same colony, nestmates, and reject other conspecific, non-nestmate [Holldobler & Wilson, 1990]. This ability is a pre-requisite and one of the conditions favoring the advent of eusociality. Nestmate recognition manifests itself at the individual level by an aggressive response to non-nestmate, either biting or dragging them outside the colony territory. Many studies quantify aggression by the presence or absence of the previous responses when two individuals are interacting, denoting an increment in antagonistic behavior when individuals transition from antennation to biting and then to dragging. Colony identity is mediated by chemicals presents on the ant's cuticles. These chemicals consist on a blend of hydrocarbons. So far, evidence is contradictory about how ants perform nestmate recognition. However, data suggests that even a single chemical can trigger a behavioral response [Guerrieri *et al*, 2009; Martin *et al*, (2008)].

As important as nestmate recognition is the ability of ants to communicate complex messages to their nestmates through the use of chemical signals acting at a distance. Nonetheless, it is difficult to discuss these behaviors without referring to the various chemicals and glands producing these chemicals, evoking behavioral responses when smelled or tasted by other colony members.

### *Glands and Pheromones*

Ants have a wide variety of different exocrine glands which major social role is the production of chemical social signals [Holldobler and Wilson, 1990]. These social cues can be divided among two broad classes: the first class, encompassing pheromones, produces action at a distance, signaling danger or indicating the path to a rich food source; the second class, containing cuticular hydrocarbons, is used to recognized nestmates from colony invaders [Billen & Morgan, 1998]. These two systems constitute the major sources of chemical communication signals and are found in almost every described ant species. Pheromones are synthesized within insects' glands following complex regulatory biosynthetic pathways, in which fatty acid precursors are elongated and hydrocarbons formed (for a review, see [Tillman et al, 1999] and an example [Legendre et al, 2008]). These chemicals form the basis of pheromones or the cuticular hydrocarbons.

Pheromones are stored in specialized glands and emitted or spread based on individual need to call for interspecific attention. Among pheromones acting at long distances, the presence of alarm and trail pheromones is widely conserved among different species. Thus, they are distinctive chemical cues belonging to the ant genus. Alarm pheromones signal the presence of an enemy and depending on the internal state of the ant, they could cause aggregation and attack or, panic and dispersal [Blum, 1969]. Alarm pheromones are primarily produced by exocrine glands located in the mandibles and the sting (varying for different ant species). As an example, the major alarm pheromone in the Pogonomyrmex genus is 4-methyl-3- heptanone [McGurck *et al*, 1966].

Trail pheromones convey the location of food sources and trigger aggregation and follow behavior [Morgan, 2009]. They are produced in the sting and are laid as ants traverse a route. Trail pheromones might comprise single compounds or blends of different ones [Morgan, 2009]. In the Pogonomyrmex genus, trail recruitment is accomplished with 3-ethyl-2,5dimethylpyrazine but addition of 2, 5-dimethylpyrazine and trimethylpyrazine creates a more attractive blend [Holldobler *et al*, 2001]. Each type of pheromone is composed of one or just a few chemicals and, the chemicals in isolation can recapitulate the behavior in the lab. The specific behavior they trigger, the completely opposite valence of these two chemical signals, and the advantage of their off-the-shelf availability make a case for studying their representation in the brain.

Cuticular hydrocarbons are secreted by the epidermis and form a coating covering the ant's cuticle; more than 1000 cuticular hydrocarbon variants have been described among different ant families (typically containing alkanes, alkenes, and their methylated forms [Martin & Drijfhout, 2009]. Although each ant species possesses its own unique CHC pattern, there is no association between CHC profile and phylogeny. This huge diversity of species-specific chemicals makes CHC ideal candidates for nest-mate discrimination signals. The specific balance of the hydrocarbon profile is the pattern assessed by some ant species to perform nestmate recognition [Martin *et al,* 2008]. In Camponotus, it has been shown that alteration of the profile by a single hydrocarbon in individual ants is sufficient to produce aggression but, addition and not absence of a single compound triggered aggression [Guerreri *et al*, 2009]. Coating glass beads with non-nestmate (nestmate) patterns triggers (or does not trigger) aggression [Ozaki et al, 2005]. These chemicals are sensed through the ant antennae, as ants antennate each other probing the chemical profile by specialized receptors located inside antenna bristles called sensillae. In camponotus Japonicus, a specialized sensillae has been identified that hosts the particular receptors [Ozaki et al, 2005].

CHC depend on a variety of sources (such as food or temperature), for this reason this complex pattern slowly changes with time. However, colonies are able to adapt and keep track of their own profile [Sorvari *et al,* 2008]. CHC have not only been implied in nestmate recognition but also to recognize the tasks of the ant encountered in the wild (such as foragers, patrollers or nurses [Greene & Gordon, 2003] as well as the identity of the queen [Vasquez *et al*, 2008].

## A.2 Behavioral Assays

To recapitulate the behavioral response of ants to the presence of pheromones and, to probe the nestmate recognition code, we developed two behavioral assays in which ants are challenged to different stimuli (individual ants or different chemicals) within their colonies or in isolation under a microscope. We emulated the set up developed by [Branson *et al*, 2007] for Drosophila in which flies are arrange into an arena with a homogeneous floor, illuminated from the top and recorded with a camera while they perform a behavioral paradigm. In our case, the arena mimics the housing of each colony and we used this arena to reproduce the nestmate recognition paradigm. For ants with a small number of individuals within the colony and sizes around 2mm in length, entire colonies can be maintained in transparent housings of eight by three centimeters. These colonies can be presented with different stimuli, in our case the introduction of a nestmate or non-nestmate ant, and their reaction can be recorded and later scored in a high-throughput manner (Figure A1).

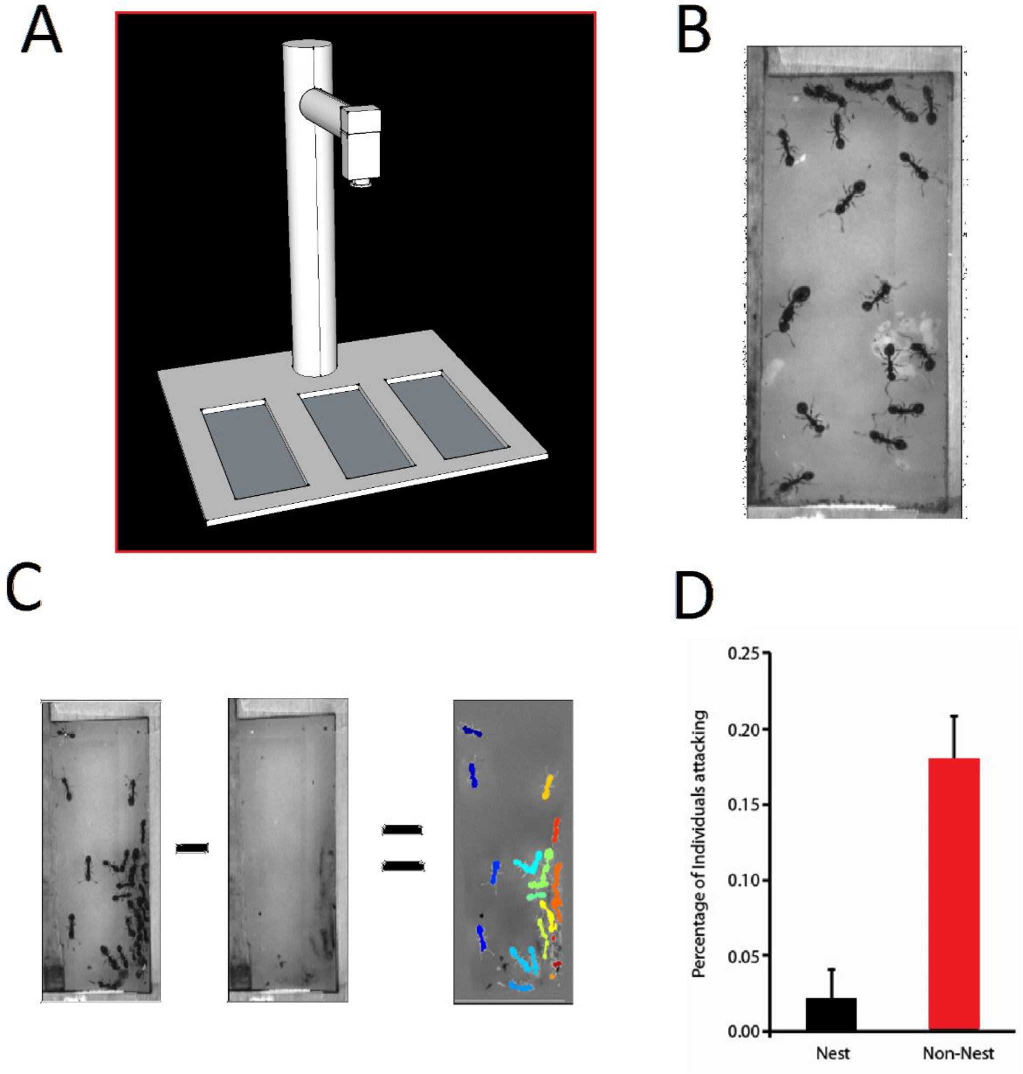


**Figure A1.1. Behavioral assay.**

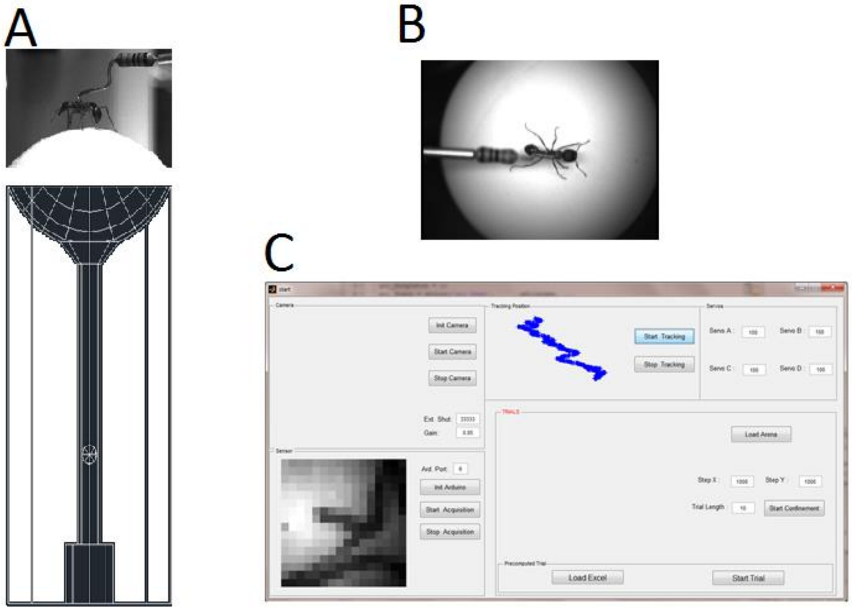
(a) Diagram representing the behavioral assay in which three colonies can be recorded simultaneously. In each of the three slots at the bottom, colonies are inserted after a stimulus is presented.

(b) Snapshot of a Temonothorax colony taken on the setup (a).

(c) Intermediate step of tracking algorithm in which background is subtracted to focus on the ants' positions.

(d) Ants can discriminate nest versus non-nestmate intruders. Colonies are randomly presented with a nestmate or non-nestmate individual and the behavior of the colony is scored for three minutes. The amount of ants attacking the intruder is shown. An ant is attacking an intruder if it is biting or dragging the introduced individual. Non intra-nest aggression was recorded in any of the experiments.

To monitor the behavior of individual ants under a microscope, we adapted the fly-on-a-ball setup developed by [Seelig *et al*, 2010] or [Kohatsu *et al*, 2011] to track the position of a tethered ant while exposed to different chemical cues. Additionally, this setup permits the monitoring of behavior while performing two-photon calcium imaging. Briefly, an ant is tethered by a miniature pin that holds them from their back and it is positioned on top of a ball floating on an air cushion. Then, the bidimensional movement of the ball is tracked using an optical sensor ADNS-3080 [ADNS3080, Avago Technologies]. We developed a matlab platform capable of recording real time video and, at the same time, this platform acquires the displacement coordinates of the ball. Through a simple mathematical transformation, these coordinates reveal the linear displacement of the animal in a fictitious environment. All this information is interfaced with an Arduino board [Arduino.org]. Code for acquiring the optical flow of the sensor is available at DIY Drones ardupilot-mega under the name AP_OpticalFlow_ADNS3080.cpp, and the full platform was informed by [Optical Flow Website].



Ext. Shut: 33333
Gain: 0.85
Servo A : 100
Servo B : 100
Servo C : 100
Servo D : 100
Ard. Port: 6
Step X : 1000
Step Y : 1000
Trial Length : 10

**Figure A1.2. Ant-on a-ball set up.**

(a) Diagram representing the holder from which a foam ball is levitated. An ant is supported at the top from a pin.

(b) The movement of the ant is monitored with a camera at the top.

(c) Matlab interface in which the movement of the ball is transformed into a two-dimensional trajectory (shown on blue).

## A.3 From the periphery into the Brain, receptors and the neural pathway processing olfactory information

*The beginning of sensory transduction*

The best studied olfactory system in the insect realm corresponds to the fruit fly, Drosophila melanogaster (for a review see, [Vosshall and Stocker, 2007]). In Drosophila, chemosensation is mediated by two sensory systems: the olfactory system, composed of olfactory receptor neurons (ORNs) located on the antennae [Vosshall et al, 2000] and, the gustatory system, composed of gustatory receptor neurons (GRNs) situated principally in the proboscis and the legs [Scott et al, 2001].

Olfactory receptors neurons project their axons through the antenna into the antennal lobe, the first relay of olfactory information in the brain located in the ventral part of the insect protocerebrum. The antennal lobe is divided into compartmentalized regions, termed glomeruli. In drosophila, a glomerulus is a structure aggregating neuropils from olfactory receptor neurons mainly expressing the same olfactory receptor [Vosshall and Stocker, 2007]. In the case of innate avoidance to CO2 and the aggressive response to the pheromone cVA, olfactory responses are mediated by a small subset of olfactory receptors (Gr21a, Gr63a and Or67d respectively) that in turn activate a small subset of antennal lobe glomeruli (V, the most ventral glomerulus and the DA1 glomerulus respectively) [Suh *et al*, 2004]; [Jones *et al*, 2007]; [Datta *et al*, 2008]; [Wang & Anderson, 2010]. Gustatory information, which is mainly non-volatile and it is mediated by contact, converge mostly to the suboesophageal ganglion. Gustatory chemosensation is responsible for the perception of appetitive food, its regulation on the extension or retraction of the proboscis and, the perception of certain Drosophila pheromones -through GRNs located in many body parts as the proboscis, wing margins, legs and ovipositor [Montell 2009]; [Thistle et al, 2012].

In other insect species, the only known olfactory receptor-ligand described is the sex pheromone (Bombykol, Bombykal) of the moth Bombyx Mori and its receptors (BmOR1). However, no thorough sampling of the full receptor repertoire has been done to guarantee this assumption [Nakagawa *et al*, 2005] [Sakurai *et al*, 2004]. In ants, olfactory receptors have been computationally annotated and their functional properties have just begun to be investigated [Zhou et al, 2012].

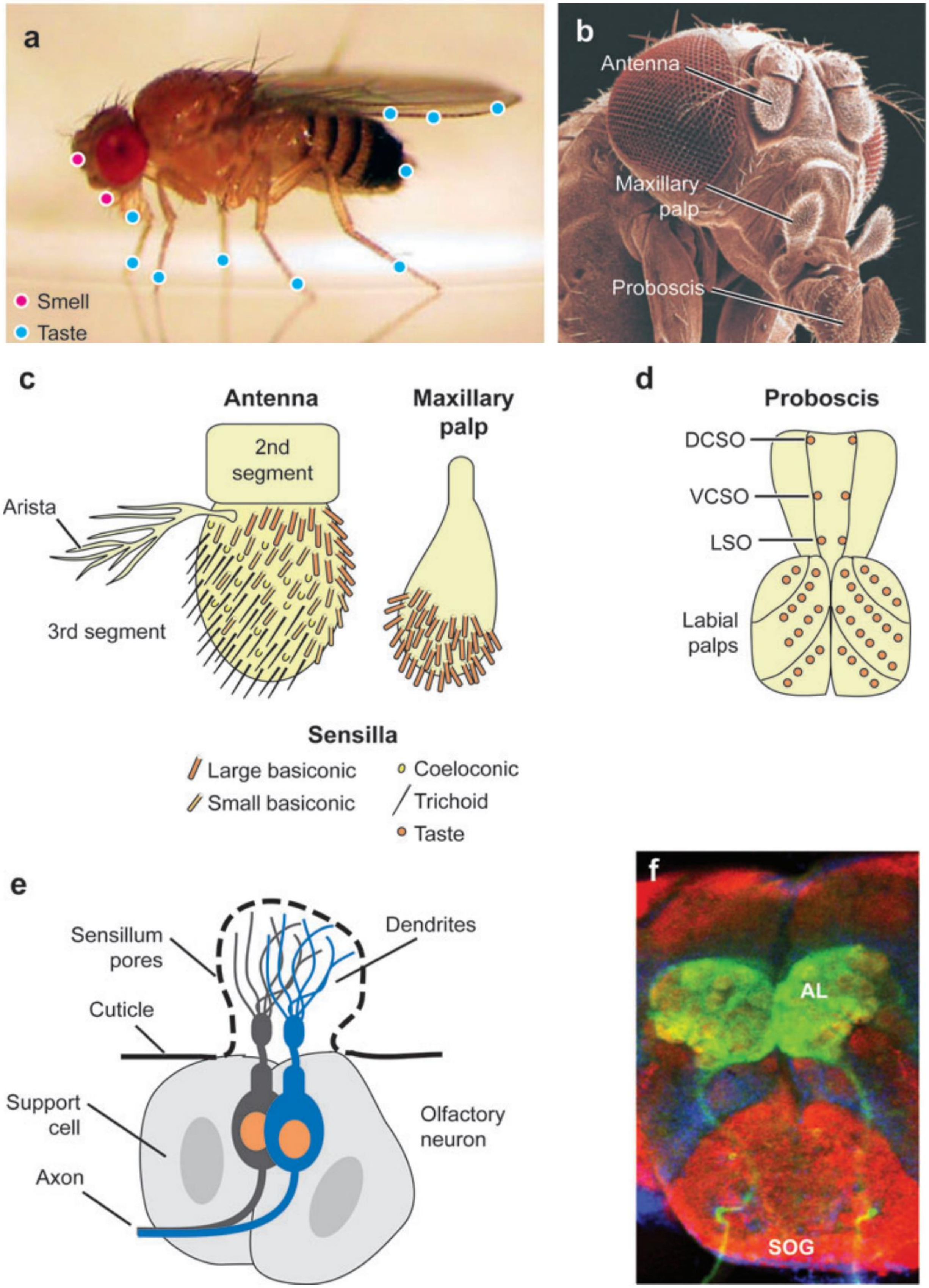

a
Smell
Taste
b
Antenna
Maxillary palp
Proboscis
c
Antenna
2nd segment
Arista
3rd segment
Maxillary palp
Sensilla
Large basiconic
Small basiconic
Coeloconic
Trichoid
Taste
d
Proboscis
DCSO
VCSO
LSO
Labial palps
e
Sensillum pores
Dendrites
Cuticle
Support cell
Olfactory neuron
Axon
f
AL
SOG

**Figure A1.3 (preceding page). Early stages of chemosensory processing in insects.**

(a) Diagram representing the location of chemosensory neurons in the fly.

(b) Scanning electron micrograph of a fly head. Image courtesy of J. Berger, MPI-Developmental Biology, Tubingen, Germany.

(c) – (d) Sensory organs indicating the position of different type of sensilla –hair like structures that host the dendrites of sensory receptor neurons.

(e) Representative olfactory sensillum in which two ORNs dendrites are housed.

(f) Antibody staining of a fly brain. On green, Or83b:GFP-labeled ORN axons, red, brain neuropil, and blue, cell nuclei.

Figure reproduced from [Vosshal and Stoker, 2007]. Panels (b-c) adapted from Benton et al. (2006), published by the Public Library of Science, which uses the Creative Commons Attribution License.

### *From the periphery into the brain*

Olfactory receptor cells send their axons into the first relay of olfactory information in the insect brain, a structure called antennal lobe. The insect antennal lobe is divided into neuropil arrangements termed glomeruli. In drosophila, it is well establish that axons from olfactory receptor cells expressing the same receptor converge to the same glomerulus and, their functional properties have been investigated by means of two photon calcium imaging [Wang *et al*, 2003]. In orthopteran (locust) a one-to-many code exists, in which olfactory receptor cells project to many micro-glomeruli. Given the lack of genetic tools, it is yet to be definitively established what model the ant antennal lobe follows. Nonetheless, given the high correlation between the number of putative olfactory receptors and the number of glomeruli, a one to one code is likely [Bonasio *et al*, 2010]; [Smith *et al*, 2011]; [Elsyk *et al*, 2016]. This indication is also repeated in another hymenoptera species as bees [HoneyBee Genome, 2006]. Even more interesting, it is hypothesized that given their single chemical representation, the antennal lobe activation by pheromones might represent a small subset of glomeruli as it has been suggested by intracellular recordings in ants [Yamagata *et al*, 2006].

From the antennal lobe, projection neurons direct their axons to the lateral horn (implicated in innate behavior), to the mushroom body calix (implicated in associated behavior) or both [Wong *et al*, 2002]; [Lin *et al*, 2007]; [Caron *et al*, 2013]; [Aso *et al*, 2014]; [Wang *et al*, 2014]. These higher order centers have been less investigated and the dual nature of the olfactory pathway and its interconnection make it difficult to deconvolve function in such an uncharted territory [Kirschner *et al*, 2006] [Galizia & Rossler 2010]; [Yamagata *et al*, 2007]. We reasoned that given the spatial segregation of olfactory information into discrete units in the antennal lobe, this region of the ant protocerebrum posed an ideal candidate region to investigate how the olfactory code differs when ant encounter friends and foes. Even more, morphological dimorphism exists between males and females in some ant species, indicating specialized antennal lobe-mediated recognition to sex-specific olfactory cues [Nishikawa *et al*, 2008]; [Mysore *et al*, 2009]; [Nakanishi *et al*, 2010]. Moreover, macroglomerulus have been implicated in differential processing of pheromones in different castes [Kuebler & Kleinedam, 2010]. These previous studies demonstrated the importance of correlating morphology to function, both fundamental dimensions of neuronal characterization that might help us understand how nestmate information is encoded. For this reason, we began the ant antennal lobe characterization by developing tools to build an atlas from which different individual functional experiments can be mapped onto.

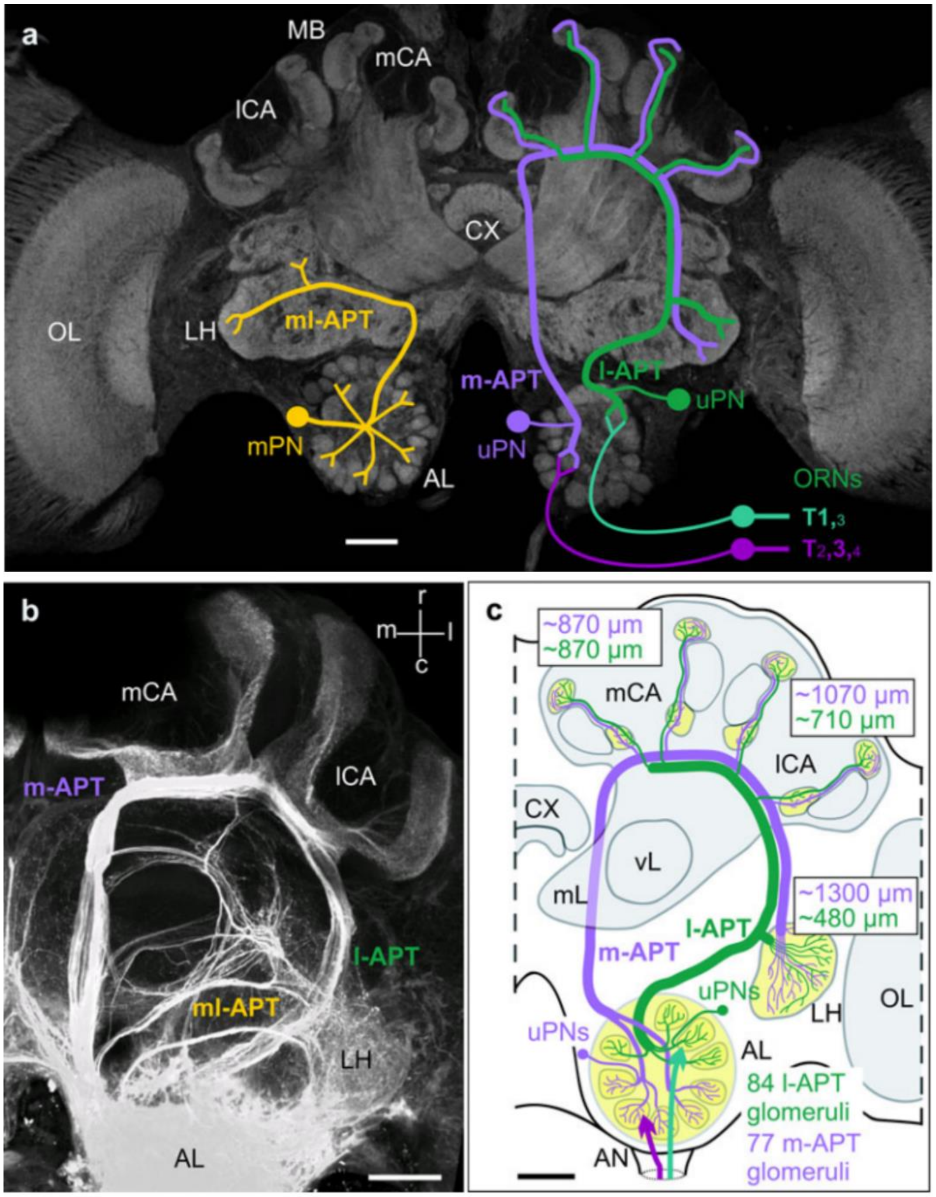


**Figure A1.4. Honey bee higher Brain Centers.**

(a) Confocal image depicting parallel olfactory systems in the honeybee brain, with a focus on the dual olfactory pathway from the antennal lobe into the lateral horn and the mushroom body.

Projection neurons (PN). Antennal-lobe protocerebral tracts (APT). Medial-tract uniglomerular PN (m-APT, uPN). Lateral-tract PN (l-APT, uPN). Multiglomerular PN (mPN). Lateral horn (LH). Medio-lateral tract (ml-APT). Adapted and modified with permission from Ro¨ssler and Zube (2011).

(b) Anterograde mass-fill of all APTs. Adapted and modified with permission from Kirschner et al. (2006).

(c) Schematic overview of the dual olfactory pathway in the honeybee.

*Computational Characterization Antennal Lobe*

With the aim of constructing a digital Atlas of the ant antennal lobe, we developed a manual pipeline to obtain confocal images of entire antennal lobes. A high contrast staining, based on [Ruchty *et al,* 2010] [see protocol A1.1] was developed to record the morphology of the ant brain even at the most ventral zones, where the microscope has to reach deep into the tissue and refraction increases and reduces the quality of the image.

**Protocol A1.1. High contrast staining of ant brain.**

- Ant brains were dissected under PBS.
- Transferred into ice-cold fixative (2% formaldehyde + 2% Glutaraldehyde in phosphate buffered saline [PBS], PH 7.2)
- Stored overnight @ 4.C.
- Brains were then rinsed in PBS (three times, 10 min).
- Brains were dehydrated in an ascending ethanol series (50, 70, 90, 95, and three times 100%, á 10 min).
- Transferred into methyl-salicylate (M-2047, Sigma-Aldrich, Steinheim, Germany).

Next, confocal images from 10 antennal lobes from the Pogonomyrmex species were segmented manually to obtain an initial dataset. These ALs consisted of 356 glomeruli (on average) ranging in size but with some enlarged glomeruli. Exploratory analysis on this dataset revealed conserved landmarks (enlarged glomeruli) that can be used to register individual antennal lobes among themselves. Brainaligner was adapted to warp ant brains among themselves [Peng *et al*, 2010]. Although subsets of glomeruli were found to be conserved, the rest were highly variable and difficult to map. This difficulty is due to the shape and size variation among glomeruli and to natural variation in glomeruli number; in drosophila, the number of glomeruli varies five percent among individuals [Jenett *et al*, 2006.]. Given this information, we expected that ~20 glomeruli might be randomly placed in each antennal lobe, a condition that introduces errors into automatic mapping software tools.

Finally, preliminary software was developed to automatically segment glomeruli based on blob segmentation [Li *et al*, 2007]; [Liu *et al*, 2008]. However, given the previously described variation, anatomy alone was considered insufficient to produce an atlas, physiology should be combined and response profiles used to uniquely identify specific glomeruli in each individual.

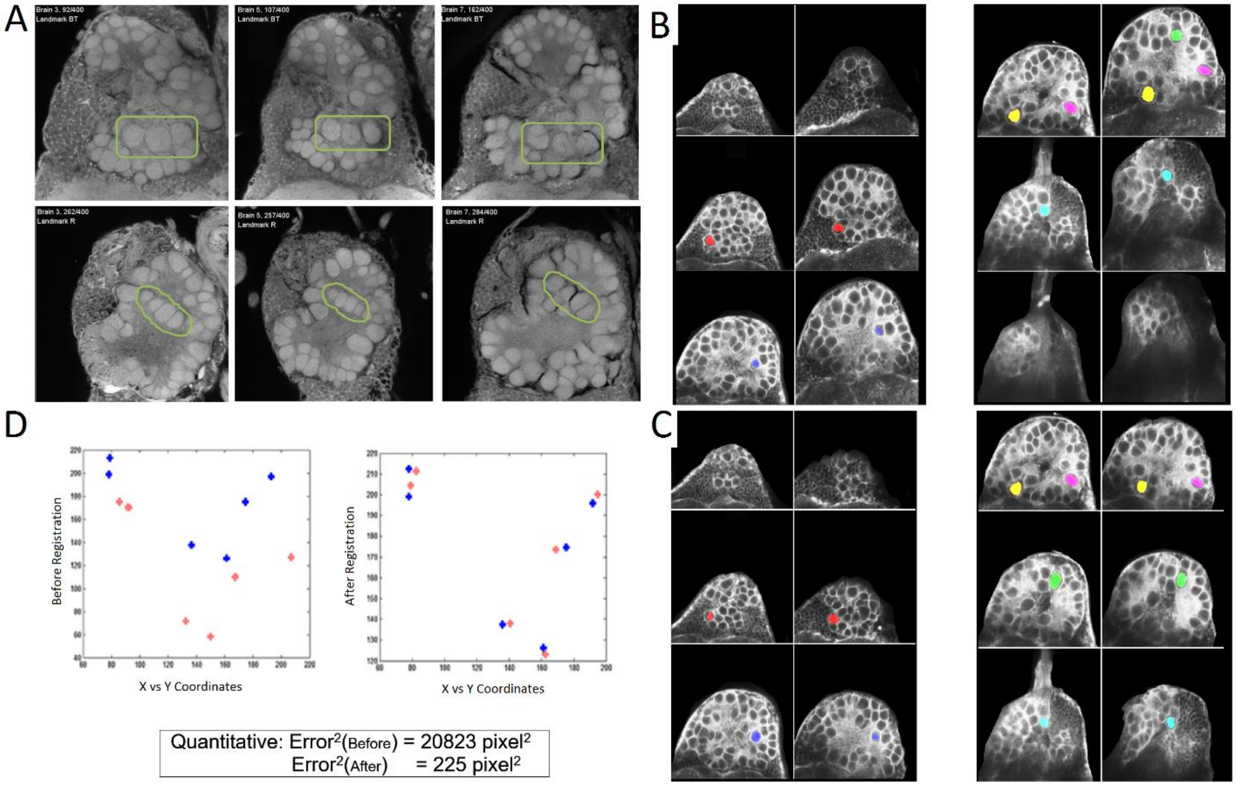


**Figure A1.5. Mapping different antennal lobes into a brain atlas.**

(a) Confocal sections of three different Pogonomyrmex brains at two different depths. Conserved set of glomeruli in each brain are encircled in green.

(b) Confocal sections of two Temnothorax brains in which landmark glomeruli are marked in red, yellow, pink and blue.

(c) Same brains as in (b) but this time the brains have been warped to a common atlas.

(d) After warping, the positional mean squared error of each landmark glomeruli is reduced.

*Calcium Imaging*

Given the anticipated complexity of the antennal lobe responses to pheromones and nestmate recognition signals [Ruchty *et al*, 2010]; [Brandstaetter & Kleineidam, 2011]; [Galizia & Menzel, 1999], I sought to develop a calcium imaging platform to sample the activity of the majority of the antennal lobe simultaneously to reveal the logic of pheromone communication in ants. To this end, I adapted an imaging preparation previously used in drosophila, ants and honey bees [Galizia *et al*, 1997]; [Want *et al*, 2000] to directly visualize the activity of glomeruli within the antennal lobe. Nonetheless, given the negative result of not visualizing clear differences between different individuals, and the difficulty in obtaining clear responses from the presentation of different chemicals, we resort to alternative methodologies to improve odor delivery and loading of calcium indicators.

## A.4 The development of a calcium imaging platform.

*Identifying optimal Stimulus Condition by Antennogram Recordings.*

Electroantennography is a biological assay that records small voltage fluctuations in whole intact insect's antennas in response to the stimulation with volatile odorants. Although a theoretical explanation is still missing, the voltage changes between the tip and the base of the antenna are believed to be caused by the depolarization of olfactory receptor neurons in synchrony [Syntech 2004]. Electroantennogram recordings were used to identify defects in peripheral expression of genes in Drosophila Melanogaster mutants [Alcorta 1991], to identify two classes of odorant pathway (in the antenna and the maxillary pulp) [Ayer & Carlson, 1992] and to identify genes in involved in olfactory signal transduction [Ayer & Carlson, 1991]. In ants, it has been previously used to optimize low volatile compound delivery methods and to establish correlations between the caste and different strength responses to pheromones in different castes [Brandstaetter *et al*, 2010]; [Kleinedam *et al*, 2005].

We used electroantennogram recordings to validate two-photon microscope preparations by recording responses during hours to different chemicals and to isolate chemicals which elicit maximal response in the antenna. We hypothesize that highly responsive chemicals might be activating many different olfactory glomeruli and this, in turn, might facilitate validation of responses on the 2p preparation. Firstly, we validated the antennogram preparation by reproducing previous measurements on drosophila melanogaster. Next, we recorded depolarization in the antenna of pogonomyrmex barbatus and identified Acetone and Ethyl acetate as chemical producing high activation. These general activation smells (in combination with a random and broad palette of smells) were used to troubleshoot the two photon microscope preparation.

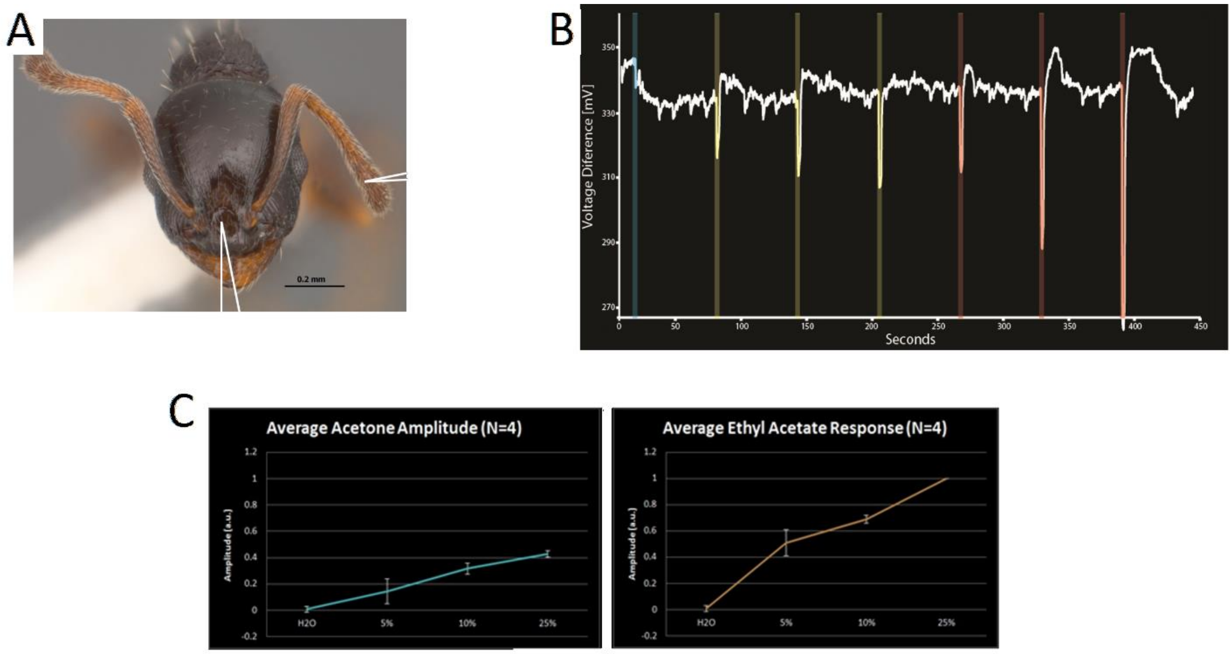


**Figure A1.6. Electroantenogram recordings.**

(a) Schematic representing the location at which recording electrodes contact the ant antenna. Ground electrode is inserted at the base of the antenna.
(b) Electroantennogram recordings depicting the presentation of two stimuli at different concentration. Yellow, acetone; red, Ethyl acetate.
(c) Average response to three different concentrations of stimuli in (b) after 5 trials.

*Improving Calcium Imaging Responses by delivering genetically encoded calcium indicators through viruses.*

Next, we sought to improve the signal to noise ratio in our two photon calcium imaging platform by delivering a genetically encoded calcium indicators (GECI) into the ant brain. A possible method to express exogenous genes into insect cells and embryos is viral delivery. Successful attempts have been made to deliver reporter genes (gfp) into drosophila S2 cells, silkworm embryos and adults and honeybee queens through the use of a baculoviruses or nucleopolyhedroviruses [Yamao *et al*, 1999]; [Kim *et al*, 2008]; [Ando *et al*, 2007]; [Ikeda *et al*, 2011].

Viruses that infect insect hosts provide ideal candidates to produce transgenic insects able to express genetically encoded calcium indicators. Candidate viruses belong to the family of Baculoviridae, and to the family of alpha viruses, like Sindbis virus. Both of these viruses are hosted by insects; silkworm in the case of Baculovirus; mosquitoes in the case of Sindbis. In addition, glycoprotein deleted rabies virus present an ideal option when their endogenous glycoprotein is replaced by the vesicular stomatitis viral glycoprotein, given the expanded host capabilities conferred by this enveloped proteins. In the case of Sindbis and Rabies virus, these two are rna-viruses, skipping the need of a promoter to drive their expression and high jacking the host cell machinery to translate their genes. Baculovirus is a DNA virus; in a typical expression system, exogenous genes are driven by an intrinsic viral promoter. However, we are not a priori certain about the functionality of this promoter in ant cells. For this reason, ant functional promoters were isolated. Describing the procedure by which these regulatory DNA sequences were identified is the goal of the next section.

Baculovirus was ultraconcentrated to achieve excessively high titer according to the manufacturer specification and subsequently purified through size exclusion chromatography [Transfiguracion *et al*, 2007]. In the same manner, rabies virus was pseudotyped and concentrated according to [Wickersham *et al*, 2010]; [Wickersham *et al*, 2013] and this method was used to concentrate Sindbis. Sindbis generation was performed according to [Foy *et al*, 2004] and [Lundstrom *et al*, 2012] (I am thankful to Thomas Reardon and Andy Murray from the Jessell laboratory, Kei Saotome and Sasha Sobolevsky from the Sobolevsky laboratory and Keneth Olson from Colorado State University for providing reagents and help in generating different viral variants). The examination of the proposed viruses is done in subsequent sections after we developed an ant cell culture assay.

*Promoter Region Identification*

A possible alternative for expressing exogenous genes in insect tissue is the identification of widely expressed promoter regions and the development of expression vectors. Efficient expression vectors have been used in the past to select transgenic embryos, to introduce foreign DNA into cells and to drive expression in electroporation experiments [Huynh & Zieler, 1999]. With the purpose of developing expression vectors useful for driving the expression of GECI in ants, we isolated different ant promoter regions from Pogonomyrmex Barbatus and tested them in cell culture.

We focused our efforts in isolating promoter regions for widely expressed genes such as Actin, Tubulin and a panneuronal marker termed Elav [O'Donnell *et al*, 1994]; [Natzle *et al*, 1984]; [O'Donnell & Wensick, 1994]; [Chung & Keller, 1990]; [Yao & White, 1994]. Using the existing genome assembly for Pogonomyrmex Barbatus, DNA regions spanning each gene was identified and the start codon mapped into the PBar genome. Based on the known promoter length for homologs genes in Dmel, candidate promoter regions were cloned and fused [Hobert 2002] to a reporter (gfp). Each DNA was introduced into two plasmid vectors, one necessary to construct a Baculovirus virus and the other one a common vector used to drive expression in Dmel insect cells.

Next, we tested such construct by transfecting them into S9 Schneider cells derived from drosophila embryos. Although ant cultured cells have not been developed, S9 cells provide a quick platform to validate our plasmid vectors for expression of the gfp reporter. Plasmids were introduced into S9 cells by using cellfectin according to the manufacturer instructions (minimal optimization of transfecting conditions was done using a positive control carrying the DMel actin

5c promoter). Tubulin and Actin resulted in the highest expression levels compared to Elav, although this result can only be stated qualitatively given the fact that no normalizing control was co-introduced with each plasmid. However, given its size (~1Kb), the tubulin promoter resulted ideal for expression system.

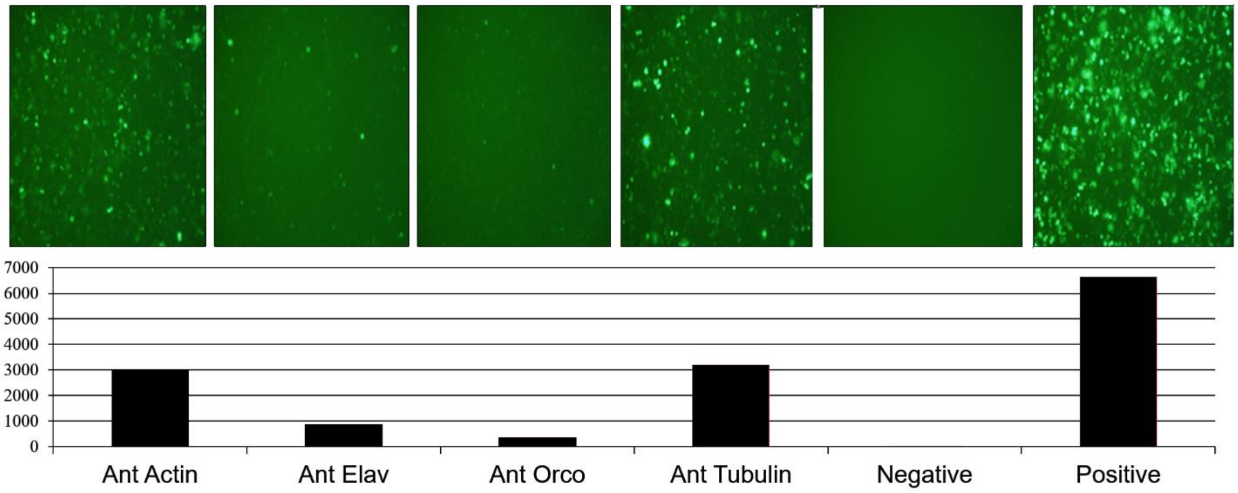


**Figure A1.7. Qualitative assessment of transfected constructs carrying ant promoters into S9 cells.**

Constructs carrying different ant promoters fused with eGFP at putative ATG positions. Transfection is performed using cellfectin reagent. Transfection conditions (not shown) were optimized using positive control (construct carrying Drosophila actin promoter expressing eGFP). It is worth noting that to perform a quantitative assessment of expression levels, a second channel should be include in which positive control would have been transfected in each condition. In this way, at every condition, a ratiometric measurement would have been performed.

(Top) Representative field of view (FOV) showing on green, gfp expressing cells.

(Bottom) Mean fluorescent of the FOV shown on Top (a.u.).

**Table A1.1. Primers used to clone ant promoter regions.**

| | |
|---|---|
| **Actin5c** | |
| Forward primer | ACCTTATTTGGTAGACAAGGTCG |
| Reverse primer | TACGAGCGCGGCAACTTC |
| Product length | 4327 |
| | |
| **Alpha Tubulin** | |
| Choice 1 | |
| Forward primer | TGGAGAGAAATGACTCGACCG |
| Reverse primer | GGCTTGTCCAACGTGGATTG |
| Product length | 1017 |
| | |
| **OrCo** | |
| Forward primer | CCATGCAGACGGCATAAACG |
| Reverse primer | ACACTTTATGCAAGTATTTGGACG |
| Product length | 3070 |
| | |
| **Elav** | |
| Forward primer | ATTTCCCCTTCTGTTCCGGG |
| Reverse primer | ACGACTGTGTCCATTCCGTT |
| Product length | 5387 |

### *Ant Cell Culture Assay*

Cell culture systems are an excellent platform to study gene expression, widely used in molecular biology, genetic and biochemical studies [Bayne 1998]; [Barbara *et al*, 2008]; [Egger *et al*, 2013]; [Kreissl & Bicker, 1992]. We developed an ant cell culture with the purpose of testing the previously developed viruses and validate their efficacy in infecting ant cells. Given our limited availability of ant embryos, ant adult brain cells were dissociated according to protocol A1.2 and then plated. Different substrates (laminin, gelatin, glass and, plastic) were tested to guarantee optimal cell adhesion, with laminin resulting in better survival. Cell culture media was adapted from [Hunter, 2010], table A1.1 and guaranteed an average survival rate of 10 days.

**Protocol A1.2. Ant brain cell dissociation protocol.**

- Adult individuals from Pogonomyrmex colonies were used. For easiness, ants were glued to individual petridishes and around ~50 individuals are used each time.
- In a sterile, laminar flow hood, ants were surface sterilized with 70% ethanol.
- Samples were then rinsed three times with PBS.
- Sterile forceps were used to dissect the brain under cell medium.
- Brains were incubated in collagenase for 3 min. Next, collagenase was inactivated by addition of new medium (containing FBS).
- Next, brains were torn apart in medium by pipetting and then dispersed across multi-well plates, 24 wells (Costar®, Corning,NY) and incubates at 25°C temperature.
- After cells displayed attachment, usually within 1d, half media was exchanged at intervals of 2d.

**Table A1.1. Cell medium composition adapted from Honey bee, A. mellifera [Hunter, 2010]**

| | |
|---|---|
| Schneider’s Insect Medium | 150 ml |
| 0.06 ML-histidine solution (pH 6.35) | 200 ml |
| Fetal Bovine Serum (heat inactivated, 56°C for 30 min) | 50 ml |
| CMRL 1066 | 15 ml |
| Hanks’ Salts | 5 ml |
| Insect medium supplement (×10) | 2 ml |
| Gentamicin, units/1 µl/ml 1 µl/ml | |
| Total volume of medium | ~422 ml; |
| pH adjusted to 6.3–6.5 with 2 N HCl or NaOH | |

Finally, we tested different viruses in concentrations similar to the ones used to infect animals ($10^7$ virions) and observed that baculovirus and rabies viruses are able to infect ant cell cultures. Experiment performed on adult individuals result in negative expression.

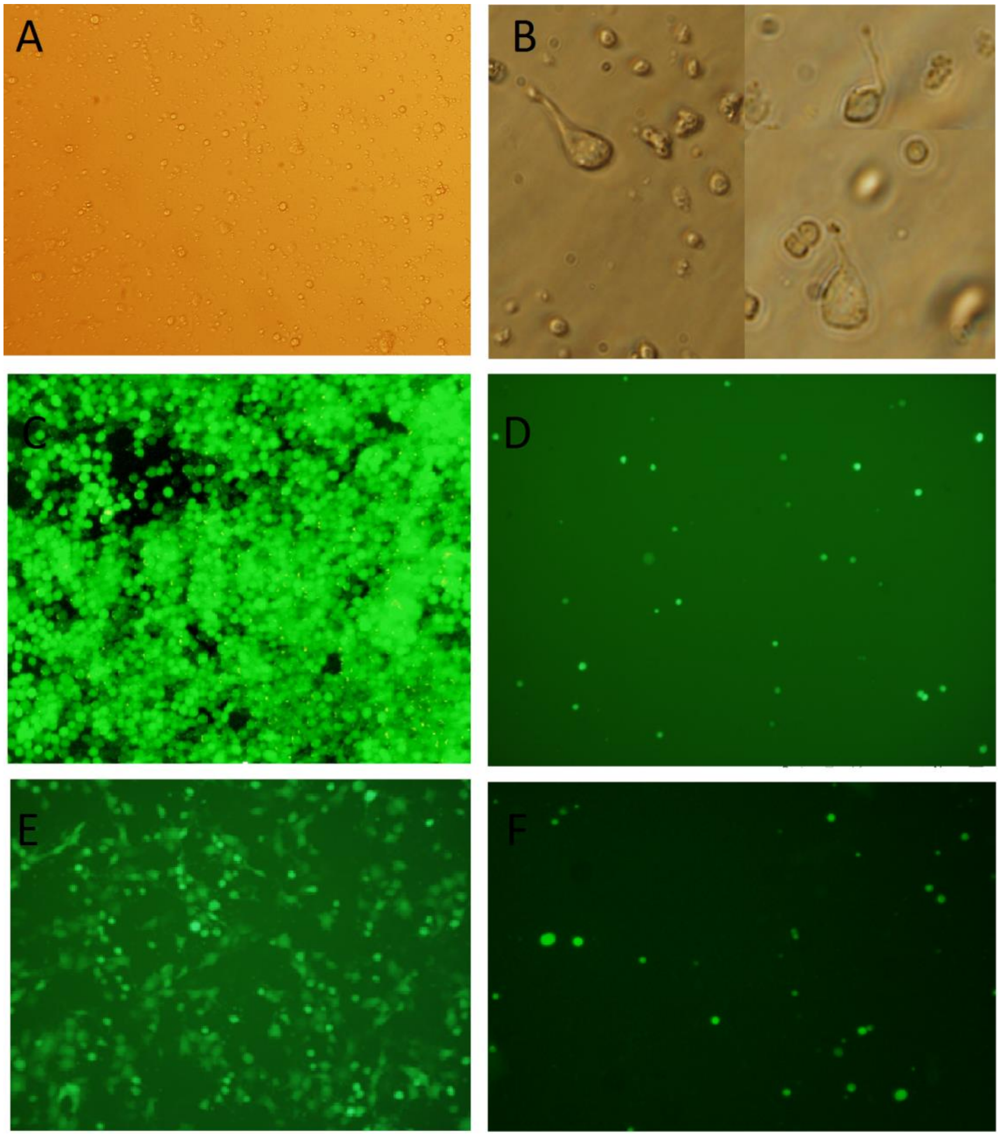


**Figure A1.8. Qualitative assessment of viral infection of ant brain cells.**

(a) Bright field depicting representative ant cell culture after 5 days.

(b) Zoom in FOV depicting ant cells growing processes.

(c) SF9 cells reporting in green infection with Baculovirus expressing GFP using the ant tubuling promoter.

(d) Same virus as in (c) infecting ant cells after 5 days. Only 10% of the cells express GFP.

(e) BHK cells reporting in green infection with pseudo-typed rabies virus carrying the VSV glycoprotein, expressing GFP.

(f) Same virus as in (e) infecting ant cells after 5 days. Only 10% of the cells express GFP.

# Appendix A - Bibliography

ADNS-3080 Avago Technologies. ADNS-3080 High-performance Optical Mouse Sensor.

Alcorta E. 1991. Characterization of the electroantenogramm in Drosophila melanogaster and its usefor identifying olfactory capture and transduction mutants. Journal of Neurophysiology 65, pp 702- 714.

Ando T., Fujiyuki T., Kawashima T., Morioka M., Kubo T., Fujiwara H. 2007. In vivo gene transfer into the honeybee using a nucleopolyhedrovirus vector. Biological and Biophysical Research Communications 352, pp 335-340.

Aso Y., Hattori D., Yu Y., Johnston R.M., Iyer N.A., Ngo T.T., Dionne H., Abbott L.F., Axel R., Tanimoto H., Rubin G.M.. 2014. The neuronal architecture of the mushroom body provides a logic for associative learning. Elife 23;3:e04577.

Ayer R.K., Carlson J. 1992. Olfactory physiology in the drosophila antenna and maxillary palp: acj6 distinguishes two classes of odorant pathways. Journal of Neurobiology 23, pp 965- 982.

Ayer R.K., Carlson J. 1991. Acj6: a gene affecting alfactory physiology and behavior in drosophila. PNAS 88, pp 5467-5471.

Bayne C.J. 1998. Methods in Cell Biology. Chp 10 Invertebrate Cell Culture considerations: Insects, Shellfish and worms. Academic Press.

Barbara G.S., Grunewald B., Paute S., Gauthier M., Raymond-Delpech V. 2008. Study of nicotinic acetylcholine receptors on cultured antennal lobe neurons from adult honeybee brains. Invertebrate Neuroscience 8, pp 19-29.

Brandstaetter A.S., Rossler W., Kleinedam C.J. 2010. Dummies versus air puffs: efficient stimulus delivery for low volatile odors. Chemical Senses 35, pp 323- 333.

Brandstaetter A.S., Kleineidam C.J. 2011. Journal of Neurophysiology 206, pp. 2437- 2449.

Billen J., Morgan D. 1998. Pheromone communication in social insects: Sources and secretions. Chapter 1. Westview Press.

Bonasio R., Zhang G., Ye C., Mutti N.S., Fang X., Qin N., Donahue G., Yang P., Li Q., Li C., Zhang P., Huang Z., Berger S.L., Reinberg D., Wang J., Liebig J. 2010. Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. Science 329, pp. 1068-1071.

Bransom K., Robie A.A., Bender J., Perona P., Dickinson M. 2007. High-throughput ethomics in large groups of Drosophila. Behavioral Ecology, pp 441- 447.

Blum M. 1969. Alarm Pheromones. Annual review of Entomology 14, pp 47-80.

Caron S.J., Ruta V., Abbott L.F., Axel R. 2013. Random convergence of olfactory inputs in the Drosophila mushroom body. Nature 497, pp. 113-117.

Chung Y.T. Keller E.B. 1990. Positive and negative regulatory elements mediating transcription from the drosophila melanogaster actin 5c distal promoter. Molecular Cellular Biology 10 (12), pp 6172-6180.

Datta S.R., Vasconcelos M.L., Ruta V., Luo S., Wong A., Demir E., Flores J., Balonze K., Dickson B.J., Axel R. 2008. The Drosophila pheromone cVA activates a sexually dimorphic neural circuit. Nature 27, pp 473-477.

Elsik C.G., Tayal A., Diesh C.M., Unni D.R., Emery M.L., Nguyen H.N., Hagen D.E. 2016. Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. Nucleic Acids Res. 44, pp. 793-800.

Egger B., van Giesen L., Moraru M., Sprecher S.G. 2013. In vitro imaging of primary neural cell culture from drosophila. Nature Protocols 8, pp 958-965.

Foy B.D., Myles K.M., Pierro D.J., Sanchez-Vargas I., Uhlirova M., Jindra M., Beaty B.J. and Olson K. 2004. Development of a new sindbis virus transducing system and its characterization in three culicine mosquitoes and two lepidopteran species. Insect Molecular Biology 13, pp 89-100.

Galizia C.G., Joerges J., Kuttner A., Faber T., Menzel R. 1997. A semi-in-vivo preparation for optical recording of the insect brain. Camponotus Rufipes. Journal of neuroscience methods 76, pp 61-69.

Galizia C.G., Sachse S., Rappert A., Menzel R. 1999. The glomerular code for odor representation is species specific in the honeybee Apis Mellifera. Nature 2, p 473-478.

Galizia C.G., Rossler W. 2010. Parallel Olfactory systems in insects: Anatomy and Function. The annual review of entomology 55, pp 399-420.

Greene M.J., Gordon D.M. 2003. Cuticular hydrocarbons inform task decisions. Nature 423, pp 32.

Guerrieri F.J., Nehring V., Jorgensen C.G., Nielsen J., Galizia C.G., d'Ettorre P. 2009. Ants recognize foes and not friends. Proceedings of the Royal Society B 282, pp 1-8.

Hobert O. 2002. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic C. elegans. Biotechniques 32, pp 728-730.

Holldobler B., Wilson E. O. 1990. The Ants. Belknap (Harvard University Press), Cambridge, MA.

Holldobler B., Morgan D.E., Oldham N.J., Liebig J. 2001. Recruitment pheromone in the harvester ant genus Pogonomyrmex. Journal of insect physiology 47, pp 369-374.

Holldobler B., Wilson E. O. 2008. The Superorganism. W. W. Norton, Publisher.

Honey Bee Genome. 2006. The Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee Apis mellifera.

Hunter W.B. 2010. Medium for development of bee cell cultures. In vitro Cellular Developmental Biology 46, pp 83-86.

Huynh C.Q. Zieler H. 1999. Construction of Modular and Versatile Plasmid vectors for the high-level expression of single or multiple genes in insects and insect cell lines. Journal Molecullar Biology 288, pp 13-20.

Ikeda T., Nakamura J., Furukawa S., Chantawannakul P., Sasaki M., Sasaki T. 2011. Transduction of baculovirus vectors to queen honeybees, Apis Mellifera. Apidologie 42, pp 461-471.

Jones, W.D., Cayirlioglu, P., Kadow, I.G., Vosshall, L.B. 2007. Two chemosensory receptors together mediate carbon dioxide detection in Drosophila. Nature 445, pp. 86-90.

Jenett A., Schindelin J., Heisenberg M. 2006. The virtual insect brain protocol: creating and comparing standardized neuroanatomy. BMC Bioinformatics 7, pp 1-12.

Jouni Sorvari, Pascal Theodora, Stefano Turillazzi, Harri Hakkarainen and Liselotte Sundström. 2007. Food resources, chemical signaling, and nest mate recognition in the ant Formica Aquilona. Behavioral Ecology, pp 441- 447.

Kirschner S., Kleinedam C.J., Zube C., Rybak J., Grunewald B., Rossler W. 2006. Dual Olfactory pathway in the Honeybee, Apis Mellifera. The journal of comparative neurology 499, pp 933-952.

Kim K.R., Kim Y.K., Cha H.J. 2008. Recombinant baculovirus-based multiple protein expression platform for Drosophila S2 cell culture. Journal of Biotechnology 133, pp 116-122.

Kleinedam C.J., Obermayer M., Halbich W., Rossler W. 2005. A macroglomerulus in the antennal lobe of Leaf-cutting ant workers and its possible functional significance. Chemical Senses 20, pp 383-392.

Kohatsu S., Koganezawa M., Yamamoto D. 2011. Female Contact Activates Male-Specific Interneurons that Trigger Stereotypic Courtship Behavior in Drosophila. Neuron 69, pp 498–508.

Kreissl S., Bicker G. 1992. Dissociated neurons of the pupal honeybee brain in cell culture. Journal of Neurocytology 21, 545-556.

Kuebler L.S., Kleinedam C.J. 2010. Distinct antennal lobe phenotypes in the leaf-cutting ant. The journal of comparative neurology 518, pp 352-365.

Legendre A., Miao X., Da Lage J., Wicker-Thomas C. 2008. Evolution of a desaturase involved in female pheromonal cuticular hydrocarbon biosynthesis and courtship behavior in Drosophila. Insect biochemistry and molecular biology 38, pp 244-255.

Li a., Liu T., Nie J., Guo L., Malicki J., Mara J., Holley SS.A., Xia W., Wong S.T.C. 2007. Detection of blob objects in microscopis zebrafish images based on gradient vector diffusion. Cytometry Part 71A, pp 835-845.

Lin H., Lai J.S., Chen Y., Chiang A. 2007. A map of olfactory representation in the Drosophila Mushroom Body. Cell 128, pp. 1205-1217.

Liu T., Li G., Nie J., Tarokh A., Zhou X., Guo L., Malicki J., Xia W., Wong S.T.C. 2008. An automated method for cell detection in Zebrafish. Neuroinformatics 5, pp 5-21.

Lundstrom K. 2012. Purification and concentration of alphavirus. Cold Spring Harbor Protocols.

Martin S.J., Vitikainen E., Helantera H., Drijfhout P. 2008. Chemical basis of nest-mate discrimination in the ant Formica Exsecta. Proc. of the Royal Soc. B 275, pp 1271-1278.

Martin S., Drijfhout F. 2009. A review of ant cuticular hydrocarbons. Journal of Chemical Ecology 35, pp 1151-1161.

McGurk D.J., Frost J., Eisenbraum E.J. 1966. Volatile compounds in ants: identification of 4-methyl-3-heptanone from pogonomyrmex ants. Journal of insect physiology 12, pp 1435-1441.

Montell C. 2009. A Taste of the Drosophila Gustatory Receptors. Cur. Opin. Neurobiol. 19, pp 345-353.

Morgan D.E. 2009. Trail pheromones of ants. Physiological Entomology 34, pp 1-17.

Mysore K., Subramanian K.A., Sarasij R.C., Suresh A., Shyamala B.V., VijayRahavan K., Rodruigues V. 2009. Caste and sex specific olfactory glomerular organization and brain structure in two sympatric ant species camponotus sericeus and camponotus compressus. Antropod structures and development 38, pp 485-497.

Nakagawa T., Sakurai T., Nishioka T., Touhara K. 2005. Insect sex-pheromone signals mediated by specific combinations of olfactory receptors. Science 307, pp 1638-1642.

Nakanishi A., Nishino H., Watanabe H., Yokohari F., Nishikawa M. 2010. Sex-specific antennal sensory system in the ant Camponotus japonicas: glomerular organizations of antennal lobes. Research in systems neuroscience 518, pp 2186-2201.

Natzle J.E., McCarthy B.J. 1984. Regulation of Drosophila \alpha – and \beta – Tubulin Genes during Development. Developmental Biology 104, pp 187 – 198.

Nishikawa M., Nishino H., Misaka Y., Kubota M., Tsuji E., Satoji Y., Yokohari F. 2008. Sexual dimorphism in the antennal lobe of the ant camponotus japonicus. Zoological Science, 25, pp 195-204

Optical Flow Website. http://www.bidouille.org/hack/mousecam.

Ruchty M., Helmchen F., Wehner R., Kleineidam C.J. 2010. Representation of thermal information in the antennal lobe of leaf-cutting ants. Front. Behav. Neurosci., Vol. 4, article 174.

Sakurai T., Nakagawa T., Mitsuno H., Mori H., Endo Y., Tanoue S., Yasukochi Y., Touhara K., Nishioka T. 2004. Identification and functional characterization of a sex pheromone receptor in the silkmoth bombyx mori. PNAS 101, pp 16653-16658.

Scott K., Brady R., Cravchik A., Morozov P., Rzhetsky A., Zuker C., Axel R. 2001. A chemosensory gene family encoding candidate gustatory and olfactory receptors in Drosophila. Cell 104, pp. 661-673.

Seelig J.D., Chiappe M.E., Lott G.K., Dutta A., Osborne J.E., Reiser M.B., Jayaraman V. 2010. Two-photon calcium imaging from head-fixed Drosophila during optomotor walking behavior. Nat. Methods. 7, pp 535-540.

Smith C.R., Smith C.D., Robertson H.M., Helmkampf M., Zimin A., Yandell M., Holt C., Hu H., Abouheif E., Benton R., *et al.* 2011. Draft genome of the red harvester ant Pogonomyrmex barbatus. Proc. Natl. Acad. Sci.108, pp. 5667-5672.

Syntech. 2004. Electroantennography, A practical introduction.

Suh, G.S.B., Wong, A.M., Hergarden, A.C., Wang, J.W., Simon, A.F., Benzer, S., Axel, R., Anderson, D.J. 2004. A single population of olfactory sensory neurons mediates an innate avoidance behavior in Drosophila. Nature 431, pp. 854-859.

Thistle R., Cameron P., Ghorayshi A., Dennison L., Scott K. 2012. Contact chemoreceptors mediate male-male repulsion and male-female attraction during Drosophila courtship. Cell 149, pp 1140-1151.

Tillman J.A., Seybold A.J., Jurenka R.A., Blomquist G.J. 1999. Insect pheromones -an overview of biosynthesis and endocrine regulation. Insect biochemistry and molecular biology 29, pp 481-514.

Transfiguracion J., Jorio H., Meghrous J., Jacob D., Kamen A. 2007. High yield purification of functional baculovirus vectors by size exclusion chromatography. Journal of virological methods 142, pp 21-28.

O'Donnell K.H., Chen C, Wensink P.C. 1994. Insulating DNA Directs Ubiquitous Transcription of the Drosophila Melanogaster \alpha 1 – tubulin Gene. Molecular and Cell Biology 14 (9), pp 6398 – 6408.

O'Donnell K.H., Wensink P.C. 1994. GAGA factor and TBF1 bind DNA elements that direct ubiquitous transcription of the \alpha 1-tubulin gene. Nucleic Acid Research 22 (22).

Ozaki M, Wada-Katsumata A, Fujikawa K, Iwasaki M, Yokohari F, Satoji Y, Nisimura T, Yamaoka R. 2005. Ant nestmate and non-nestmate discrimination by a chemosensory sensillum. Science 309, pp 311-314.

Peng H., Chung P., Long F., Jenett A., Seeds A.M., Myers E.W., Simpson J.H. 2010. BrainALigner: 3D registration atlases of Drosophila brains. Nature Methods 8, pp 493-500.

Vasquez G.M., Schal C., Silverman J. 2008. Cuticular hydrocarbons as queen adoption cues in the invasive argentine ant. The journal of experimental biology 211, pp 1249-1256.

Vosshall L.B., Wong A.M., Axel R. 2000. An olfactory sensory map in the fly brain. Cell 102, pp 147-159.

Vosshall L.B., Stocker R.F. 2007. Molecular Architecture of Smell and Taste in Drosophila. Annual Review of Neuroscience 30, pp. 505-533.

Wang L., Anderson D. 2010. Identification of an aggression-promoting pheromone and its receptor neurons in Drosophila. Nature 463, pp 227-232.

Wang J.W., Wong A.M., Flores J., Vosshall L.B., Axel R. 2003. Two-photon calcium imaging reveals an odor-evoked map of activity in the fly brain. Cell 112, pp 271-282.

Wang K., Gong J., Wang Q., Li H., Cheng Q., Liu Y., Zeng S., Wang Z. 2014. Parallel pathways convey olfactory information with opposite polarities in Drosophila. Proc. Natl. Acad. Sci. USA 111, pp 3164-3169.

Wickersham I., Sullivan H.A., Seung S.H. 2010. Production of glycoprotein-deleted rabies viruses for monosynaptic tracing and high-level gene expression in neurons. Nature Protocols 5, pp 595-606.

Wickersham I.R., Sullivan H.A., Seung H.S. 2013. Axonal and subcellular labelling using modified rabies viral vectors. Nature Communication 4, pp 2332-2340.

Wong AM, Wang JW, Axel R. 2002. Spatial representation of the glomerular map in the Drosophila protocerebrum. Cell 109, pp 229-241.

Yamagata N, Nishino H, Mizunami M. 2006. Pheromone-sensitive glomeruli in the primary olfactory centre of ants. Proceedings Biology Science 273, pp. 2219-2225.

Yamagata N, Nishino H, Mizunami M. 2007. Neural pathways for the processing of alarm pheromone in the ant brain. J Comp Neurol. 505, pp. 424-442.

Yamao M., Katayama N., Nakazawa H., Hayashi Y., Hara S., Kamei K., Mori H. 1999. Gene targeting in the silkworm by use of a baculovirus. Genes & Development 13, pp 511-516.

Yao K., White L. 1994. Neural Specificity of elav expression: defining a Drosophila promoter for directing expression to the nervous system. Journal of Neurochemistry 63, pp 41-51.

Zhou X., Slone J.D., Rokas A., Berger S.L., Liebig J., Ray A., Reinberg D., Zwiebel L.J. 2012. Phylogenetic and transcriptomic analysis of chemosensory receptors in a pair of divergent ant species reveals sex-specific signatures of odor coding. PLoS Genetics 8, e1002930.

# Appendix A

# Studying the neural substrate of nestmate recognition in ants.

## A.1 Introduction

This section is presented as reference for future studies. Ants live in one of the most amazing complex societies from the animal kingdom, achieving one of the highest levels of organization, termed eusociality (greek: eu “truly”, social). Their lives are organized around the maintenance of the colony, sacrificing their individuality in pursuit of the group wellbeing [Holldobler & Willson, 2008]. Depending on the ant species considered, each colony can contain from ten to as many as twenty million members. Interactions among members of these superorganisms are regulated by secreted chemical cues that convey a variety of messages, from the presence of an intruder to the location of a food source. Even more, the recognition of nestmates is also communicated through an array of chemicals which reside on the ants’ cuticles. In this appendix, we investigate the development of techniques aimed to explore the neural substrate mediating the recognition of chemical signals in the ant brain.

### *Behavior*

A hallmark among ant societies is the ability of individuals to recognize members of the same colony, nestmates, and reject non-nestmate [Holldobler & Wilson, 1990]. This ability is a pre-requisite and one of the conditions favoring the advent of eusociality. Nestmate recognition manifests itself at the individual level by an aggressive response to non-nestmate, either biting or dragging them outside the colony territory. Many studies quantify aggression by the presence or absence of the previous responses when two individuals are interacting, denoting an increment in antagonistic behavior when individuals transition from antennation to biting and then to dragging. Colony identity is mediated by chemicals presents on the ant's cuticles. These chemicals consist on a blend of hydrocarbons. So far, evidence is contradictory about how ants perform nestmate recognition. However, data suggests that even a single chemical can trigger a behavioral response [Guerrieri *et al*, 2009; Martin *et al*, (2008)].

As important as nestmate recognition is the ability of ants to communicate complex messages to their nestmates through the use of chemical signals acting at a distance. It is difficult to discuss these behaviors without referring to the various chemicals and glands producing these chemicals. For this reason, in the next section we examined the many different chemicals and glands mediating chemical communication among ants.

### *Glands and Pheromones*

Ants have a wide variety of different exocrine glands that produce an array of chemical social signals [Holldobler and Wilson, 1990]. These social cues can be divided among two broad classes: the first class, encompassing pheromones, produces action at a distance, signaling danger or indicating the path to a rich food source; the second class, containing cuticular hydrocarbons, is used to recognized nestmates from colony invaders [Billen & Morgan, 1998]. These two systems constitute the major sources of chemical communication signals and are found in almost every ant species. Pheromones are synthesized within insects' glands through complex biosynthetic pathways, in which fatty acid precursors are extended and hydrocarbons formed (for a review, see [Tillman et al, 1999] and an example [Legendre et al, 2008]). These chemicals form the basis of pheromones or the cuticular hydrocarbons.

Pheromones are stored in specialized glands and emitted or spread based on individual need to call for interspecific attention. Among pheromones acting at long distances, the presence of alarm and trail pheromones is widely conserved among different species. Alarm pheromones signal the presence of an enemy and depending on the internal state of the ant, they could cause aggregation and attack or, panic and dispersal [Blum, 1969]. Alarm pheromones are primarily produced by exocrine glands located in the mandibles and the sting (varying for different ant species). As an example, the major alarm pheromone in the Pogonomyrmex genus is 4-methyl-3-heptanone [McGurck *et al*, 1966].

Trail pheromones convey the location of food sources and trigger aggregation and path following behavior [Morgan, 2009]. They are produced in the sting and are laid as ants traverse a route. Trail pheromones might comprise single compounds or blends of different ones [Morgan, 2009]. In the Pogonomyrmex genus, trail recruitment is accomplished with 3-ethyl-2,5dimethylpyrazine but addition of 2, 5-dimethylpyrazine and trimethylpyrazine creates a more attractive blend [Holldobler *et al*, 2001]. Each type of pheromone is composed of one or just a few chemicals and, the chemicals in isolation can recapitulate the behavior in the lab. The stereotyped behavior that alarm and trail signals cause, their completely opposite valence and, the advantage of their off-the-shelf availability make a case for studying their representation in the brain.

Cuticular hydrocarbons are secreted by the epidermis and form a coating covering the ant's cuticle; more than 1000 cuticular hydrocarbon variants have been described among different ant families (typically containing alkanes, alkenes, and their methylated forms [Martin & Drijfhout, 2009]. Although each ant species possesses its own unique CHC pattern, there is no association between CHC profile and phylogeny. This vast diversity of species-specific chemicals makes CHC ideal candidates for nest-mate discrimination signals. The specific balance of the hydrocarbon profile is the pattern assessed by some ant species to perform nestmate recognition [Martin *et al*, 2008]. In Camponotus, it has been shown that alteration of the profile by a single hydrocarbon in individual ants is sufficient to cause aggression [Guerreri *et al*, 2009]. Coating glass beads with non-nestmate (nestmate) patterns triggers (or does not trigger) aggression [Ozaki et al, 2005]. These chemicals are sensed through the ant antennae, as ants antennate each other probing the chemical profile by specialized receptors located inside antenna bristles called sensillae. In camponotus Japonicus, a specialized basiconica sensillae host such receptors [Ozaki et al, 2005]. Different sensilla, within the basiconica class, respond in a different way to CHCs. It has been suggested that the differential response of basiconica sensilla is used to recognize nestmate from non-nestmates [Sharma et al, 2015].

## A.2 Behavioral Assays

To recapitulate the behavioral response of ants to the presence of pheromones and, to probe the nestmate recognition code, we developed two behavioral assays in which ants are challenged to different stimuli (individual ants or different chemicals) within their colonies or in isolation under a microscope. We emulated the set up developed by [Branson *et al*, 2007] for Drosophila in which flies are arrange into an arena with a homogeneous floor, illuminated from the top and recorded with a camera while they perform a behavioral paradigm. In our case, the arena mimics the housing of each colony and we used this arena to reproduce the nestmate recognition paradigm. For ants with a small number of individuals within the colony and sizes around 2mm in length, entire colonies can be maintained in transparent housings of eight by three centimeters. These colonies can be presented with different stimuli, in our case the introduction of a nestmate or non-nestmate ant, and their reaction can be recorded and later scored in a high-throughput manner (Figure A1).

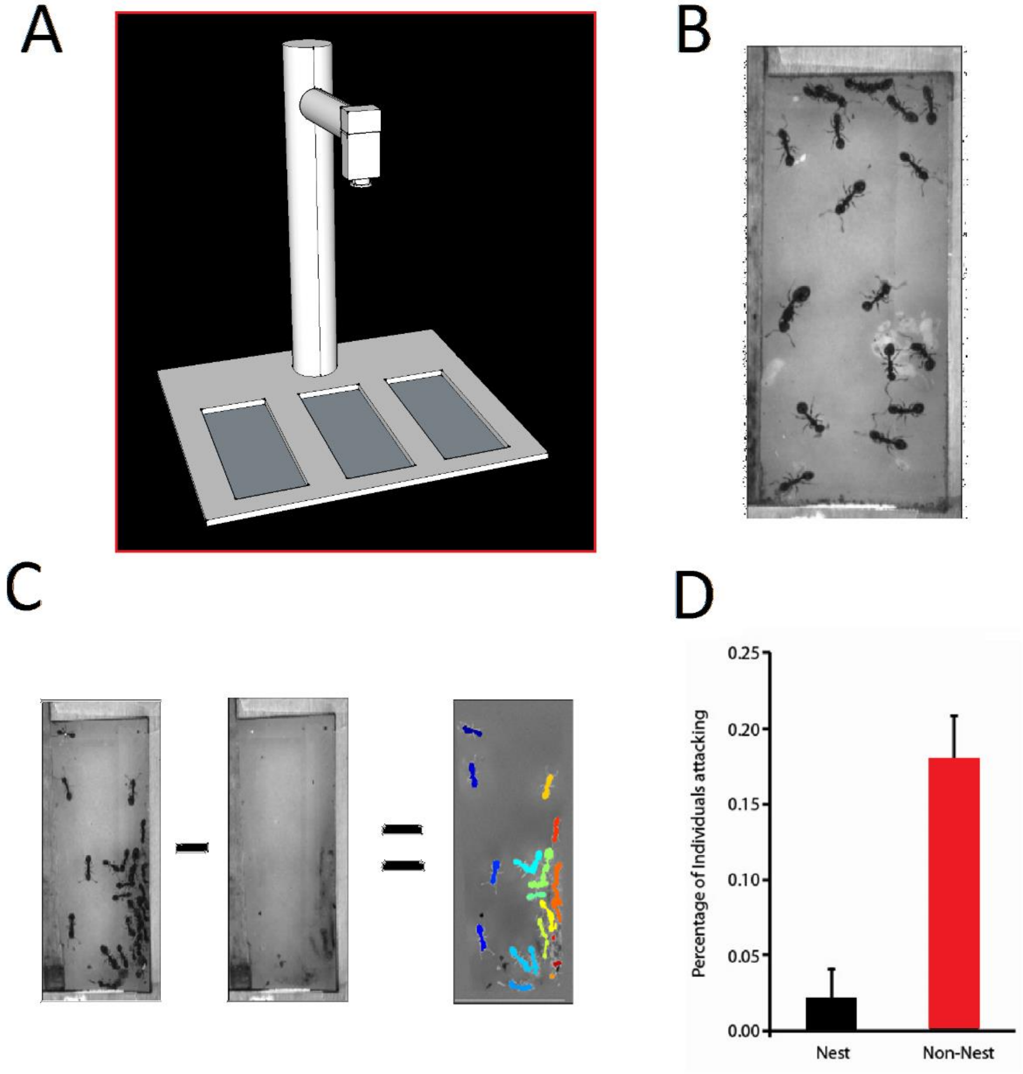


**Figure A1.1. Behavioral assay.**

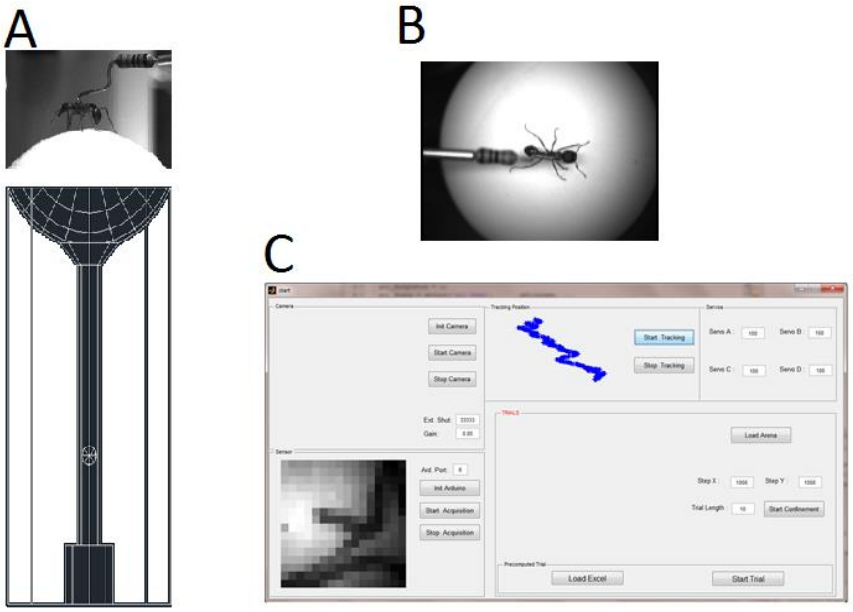
(a) Diagram representing the behavioral assay in which three colonies can be recorded simultaneously. In each of the three slots at the bottom, colonies are inserted after a stimulus is presented.

(b) Snapshot of a Temonothorax colony taken on the setup (a).

(c) Intermediate step of tracking algorithm in which background is subtracted to focus on the ants' positions.

(d) Ants can discriminate nest versus non-nestmate intruders. Colonies are randomly presented with a nestmate or non-nestmate individual and the behavior of the colony is scored for three minutes. The amount of ants attacking the intruder is shown. An ant is attacking an intruder if it is biting or dragging the introduced individual. Non intra-nest aggression was recorded in any of the experiments.

To monitor the behavior of individual ants under a microscope, we adapted the fly-on-a-ball setup developed by [Seelig *et al*, 2010] or [Kohatsu *et al*, 2011] to track the position of a tethered ant while exposed to different chemical cues. Additionally, this setup should permit the monitoring of behavior while performing two-photon calcium imaging. Briefly, an ant is tethered by a miniature pin that holds them from their back and it is positioned on top of a ball floating on an air cushion. Then, the bidimensional movement of the ball is tracked using an optical sensor ADNS-3080 [ADNS3080, Avago Technologies]. We developed a matlab platform capable of recording real time video and, at the same time, this platform acquires the displacement coordinates of the ball. Through a simple mathematical transformation, these coordinates reveal the linear displacement of the animal in a virtual environment. All this information is interfaced with an Arduino board [Arduino.org]. Code for acquiring the optical flow of the sensor is available at DIY Drones ardupilot-mega under the name AP_OpticalFlow_ADNS3080.cpp, and the full platform was informed by [Optical Flow Website].



Servo A : 100
Servo B : 100
Servo C : 100
Servo D : 100
Ext. Shut: 33333
Gain: 0.85
Ard. Port: 6
Step X : 1000
Step Y : 1000
Trial Length : 10

**Figure A1.2. Ant-on a-ball set up.**

(a) Diagram representing the holder from which a foam ball is levitated. An ant is supported at the top from a pin.

(b) The movement of the ant is monitored with a camera at the top.

(c) Matlab interface in which the movement of the ball is transformed into a two-dimensional trajectory (shown on blue).

## A.3 From the periphery into the Brain, receptors and the neural pathway processing olfactory information

### *The beginning of sensory transduction*

The best studied olfactory system in the insect realm corresponds to the fruit fly, Drosophila melanogaster (for a review see, [Vosshall and Stocker, 2007]). In Drosophila, chemosensation is mediated by two sensory systems: the olfactory system, composed of olfactory receptor neurons (ORNs) located on the antennae [Vosshall et al, 2000] and, the gustatory system, composed of gustatory receptor neurons (GRNs) situated principally in the proboscis and the legs [Scott et al, 2001].

Olfactory receptors neurons project their axons through the antenna into the antennal lobe, the first relay of olfactory information in the brain located in the ventral part of the insect protocerebrum.  The antennal lobe is divided into compartmentalized regions, termed glomeruli. In drosophila, a glomerulus is a structure aggregating neuropils from olfactory receptor neurons mainly expressing the same olfactory receptor [Vosshall and Stocker, 2007]. In the case of innate avoidance to CO2 and the aggressive response to the pheromone cVA, olfactory responses are mediated by a small subset of olfactory receptors (Gr21a, Gr63a and Or67d respectively) that in turn activate a small subset of antennal lobe glomeruli (V, the most ventral glomerulus and the DA1 glomerulus respectively) [Suh *et al*, 2004]; [Jones *et al*, 2007]; [Datta *et al*, 2008]; [Wang & Anderson, 2010]. Gustatory information, which is mainly non-volatile and it is mediated by contact, converge mostly to the suboesophageal ganglion. Gustatory chemosensation is responsible for the perception of appetitive food, its regulation on the extension or retraction of the proboscis and, the perception of certain Drosophila pheromones -through GRNs located in many body parts as the proboscis, wing margins, legs and ovipositor [Montell 2009]; [Thistle et al, 2012].

In other insect species, the only known olfactory receptor-ligand is the sex pheromone (Bombykol, Bombykal) of the moth Bombyx Mori and its receptors (BmOR1). However, no systematic sampling of the full receptor repertoire has been done to guarantee this assumption [Nakagawa *et al*, 2005] [Sakurai *et al*, 2004]. In ants, olfactory receptors have been computationally annotated and their functional properties have just begun to be investigated [Zhou et al, 2012].

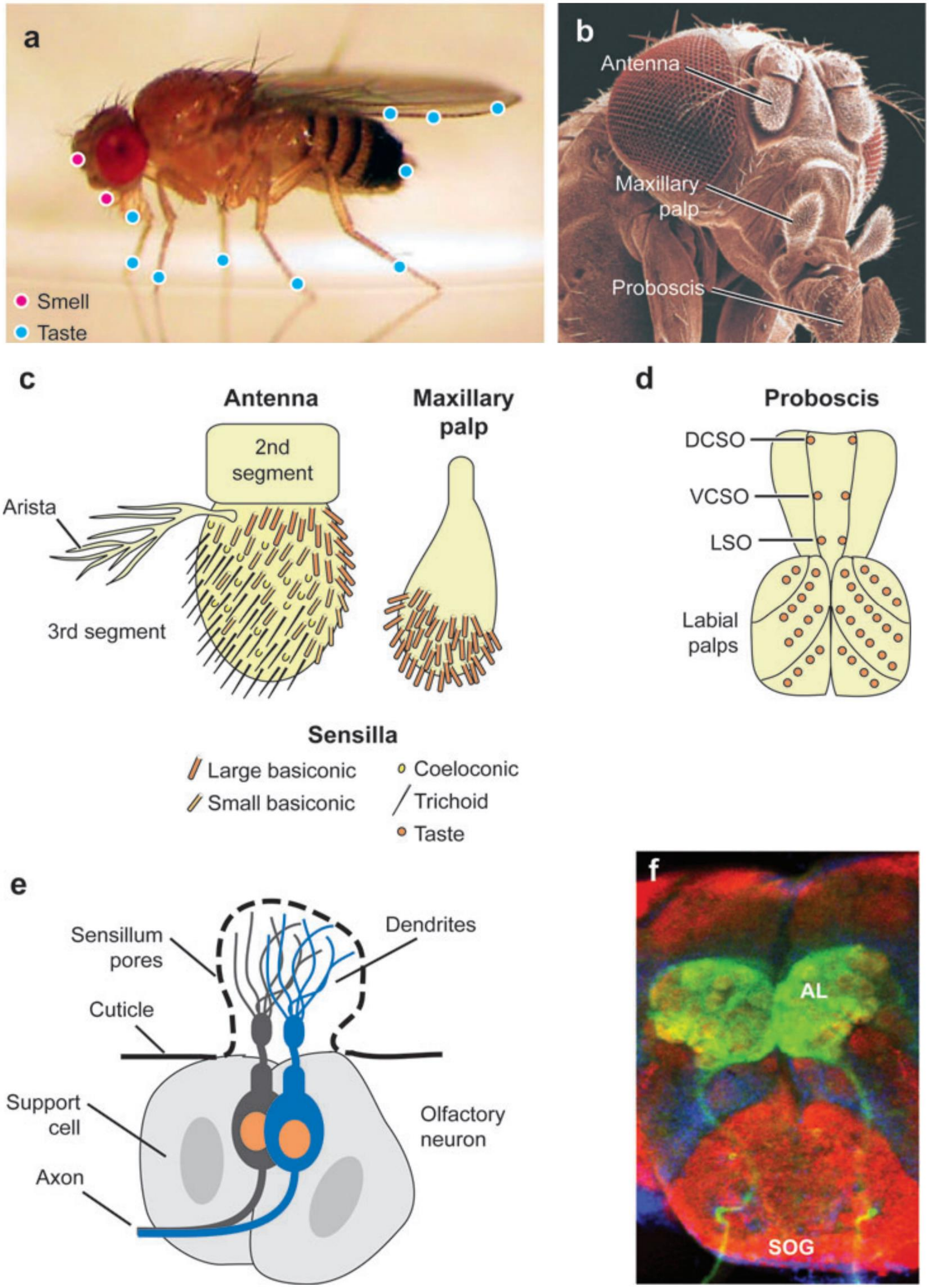

a
● Smell
● Taste
b
Antenna
Maxillary palp
Proboscis
c
Antenna
2nd segment
Arista
3rd segment
Maxillary palp
d
Proboscis
DCSO
VCSO
LSO
Labial palps
Sensilla
Large basiconic
Small basiconic
Coeloconic
Trichoid
Taste
e
Sensillum pores
Dendrites
Cuticle
Support cell
Olfactory neuron
Axon
f
AL
SOG

**Figure A1.3 (preceding page). Early stages of chemosensory processing in insects.**

(a) Diagram representing the location of chemosensory neurons in the fly.

(b) Scanning electron micrograph of a fly head. Image courtesy of J. Berger, MPI-Developmental Biology, Tubingen, Germany.

(c) – (d) Sensory organs indicating the position of different type of sensilla –hair like structures that host the dendrites of sensory receptor neurons.

(e) Representative olfactory sensillum in which two ORNs dendrites are housed.

(f) Antibody staining of a fly brain. On green, Or83b:GFP-labeled ORN axons, red, brain neuropil, and blue, cell nuclei.

Figure reproduced from [Vosshal and Stoker, 2007]. Panels (b-c) adapted from Benton et al. (2006), published by the Public Library of Science, which uses the Creative Commons Attribution License.

*From the periphery into the brain*

Olfactory receptor cells send their axons into the first relay of olfactory information in the insect brain, a structure called antennal lobe. The insect antennal lobe is divided into neuropil arrangements termed glomeruli. In drosophila, it is well establish that axons from olfactory receptor cells expressing the same receptor converge to the same glomerulus and, their functional properties have been investigated by means of two photon calcium imaging [Wang *et al*, 2003]. In orthopteran (locust) a one-to-many code exists, in which olfactory receptor cells project to many micro-glomeruli. Given the lack of genetic tools, it is yet to be definitively established what model the ant antennal lobe follows. Nonetheless, given the high correlation between the number of putative olfactory receptors and the number of glomeruli, a one to one code is likely [Bonasio *et al*, 2010]; [Smith *et al*, 2011]; [Elsyk *et al*, 2016]. This indication is also repeated in another hymenoptera species as bees [HoneyBee Genome, 2006]. However, intracellular recordings and subsequent dye fillings, performed on next order projection neurons in the antennal lobe of Camponotus, revealed that five glomeruli respond to alarm pheromone [Yamagata *et al*, 2006]. This result contradicts the one-glomeruli – one-receptor logic.

From the antennal lobe, projection neurons direct their axons to the lateral horn (implicated in innate behavior), to the mushroom body calix (implicated in associated behavior) or both [Wong *et al*, 2002]; [Lin *et al*, 2007]; [Caron *et al*, 2013]; [Aso *et al*, 2014]; [Wang *et al*, 2014]. These higher order centers have been less investigated and the dual nature of the olfactory pathway and its interconnection make it difficult to deconvolve function in such an uncharted territory [Kirschner *et al*, 2006] [Galizia & Rossler 2010]; [Yamagata *et al*, 2007]. We reasoned that given the spatial segregation of olfactory information into discrete units in the antennal lobe, this region of the ant protocerebrum posed an ideal candidate region to investigate how the olfactory code differs when ant encounter friends and foes.

Macroglomerulus in the antennal lobe have been implicated in differential processing of pheromones in different castes [Kuebler & Kleinedam, 2010]. This results suggests that morphology is an extremely helpful feature when dissecting glomeruli function. Additionally, detailed morphological information from many individuals can be used to correlate responses from different individuals. For this reason, we began the ant antennal lobe characterization by developing tools to build an atlas from which different individual functional experiments can be mapped onto.

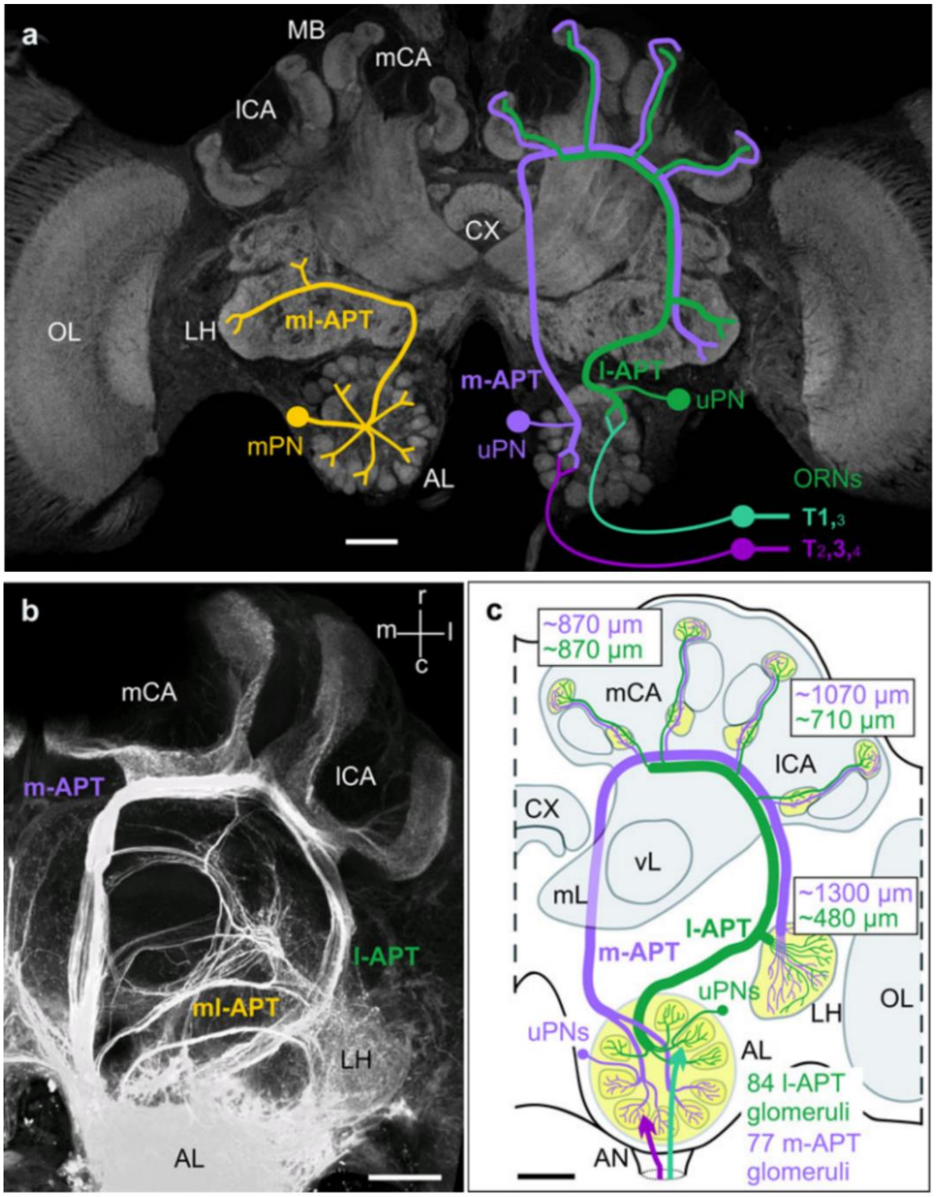


**Figure A1.4. Honey bee higher Brain Centers.**

(a) Confocal image depicting parallel olfactory systems in the honeybee brain, with a focus on the dual olfactory pathway from the antennal lobe into the lateral horn and the mushroom body.

Projection neurons (PN). Antennal-lobe protocerebral tracts (APT). Medial-tract uniglomerular PN (m-APT, uPN). Lateral-tract PN (l-APT, uPN). Multiglomerular PN (mPN). Lateral horn (LH). Medio-lateral tract (ml-APT). Adapted and modified with permission from Ro¨ssler and Zube (2011).

(b) Anterograde mass-fill of all APTs. Adapted and modified with permission from Kirschner et al. (2006).

(c) Schematic overview of the dual olfactory pathway in the honeybee.

*Computational Characterization Antennal Lobe*

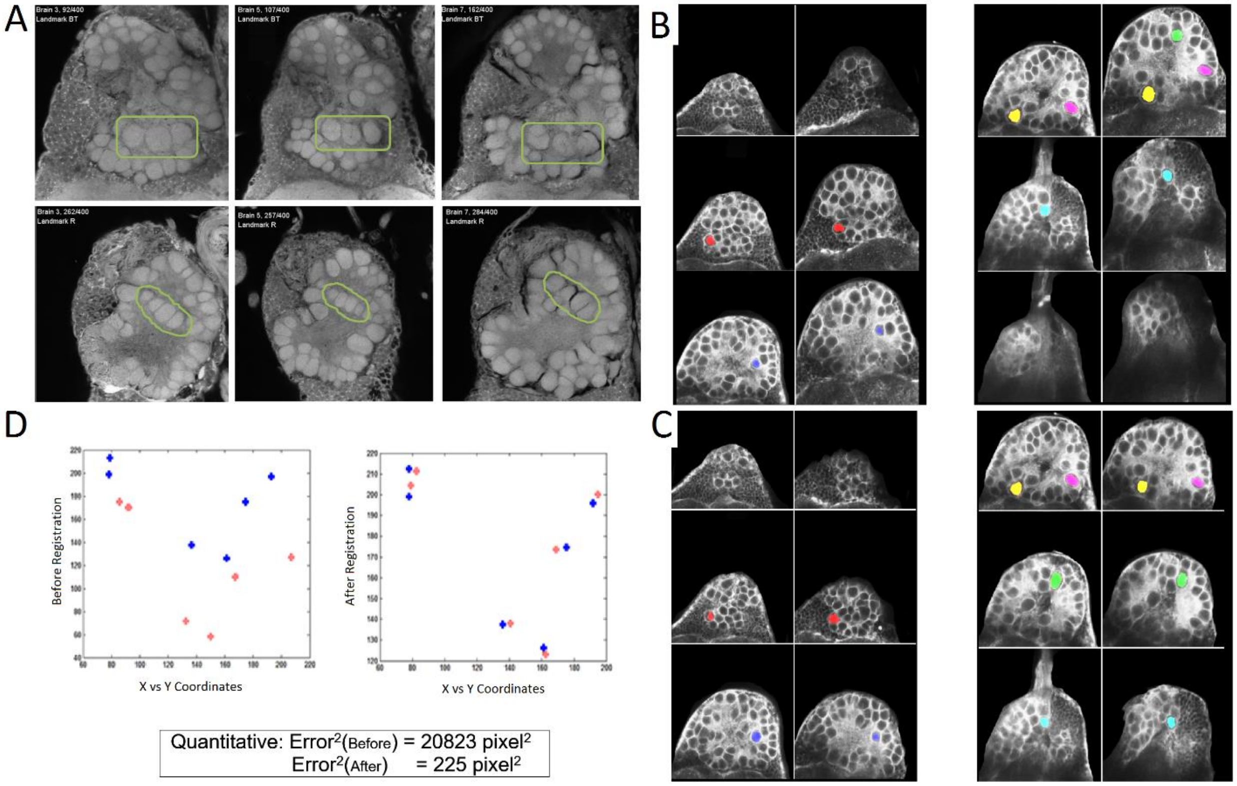
With the aim of constructing a digital Atlas of the ant antennal lobe, we developed a manual pipeline to obtain confocal images of entire antennal lobes. A high contrast staining, based on [Ruchty *et al,* 2010] [see protocol A1.1] was developed to record the morphology of the ant brain even at the most ventral zones, where the microscope has to reach deep into the tissue and refraction increases and reduces the quality of the image.

**Protocol A1.1. High contrast staining of ant brain.**

- Ant brains were dissected under PBS.
- Transferred into ice-cold fixative (2% formaldehyde + 2% Glutaraldehyde in phosphate buffered saline [PBS], PH 7.2)
- Stored overnight @ 4.C.
- Brains were then rinsed in PBS (three times, 10 min).
- Brains were dehydrated in an ascending ethanol series (50, 70, 90, 95, and three times 100%, á 10 min).
- Transferred into methyl-salicylate (M-2047, Sigma-Aldrich, Steinheim, Germany).

Next, confocal images from 10 antennal lobes from the Pogonomyrmex species were segmented manually to obtain an initial dataset. These ALs consisted of 356 glomeruli (on average) ranging in size but with some enlarged glomeruli. Exploratory analysis on this dataset revealed conserved landmarks (enlarged glomeruli) that can be used to register individual antennal lobes among themselves. Brainaligner was adapted to warp ant brains among themselves [Peng *et al*, 2010]. Although subsets of glomeruli were found to be conserved, the rest were highly variable and difficult to map. This difficulty is due to the shape and size variation among glomeruli and to natural variation in glomeruli number; in drosophila, the number of glomeruli varies five percent among individuals [Jenett *et al*, 2006.]. Given this information, we expected that ~20 glomeruli might be randomly placed in each antennal lobe, a condition that introduces errors into automatic mapping software tools.

Finally, preliminary software was developed to automatically segment glomeruli based on blob segmentation [Li *et al*, 2007]; [Liu *et al*, 2008]. However, given the previously described variation, anatomy alone was considered insufficient to produce an atlas, physiology should be combined and response profiles used to uniquely identify specific glomeruli in each individual.



Quantitative: $\text{Error}^2(\text{Before}) = 20823$ pixel$^2$
$\text{Error}^2(\text{After}) = 225$ pixel$^2$

**Figure A1.5. Mapping different antennal lobes into a brain atlas.**

(a) Confocal sections of three different Pogonomyrmex brains at two different depths. Conserved set of glomeruli in each brain are encircled in green.

(b) Confocal sections of two Temnothorax brains in which landmark glomeruli are marked in red, yellow, pink and blue.

(c) Same brains as in (b) but this time the brains have been warped to a common atlas.

(d) After warping, the positional mean squared error of each landmark glomeruli is reduced.

*Calcium Imaging*

Given the anticipated complexity of the antennal lobe responses to pheromones and nestmate recognition signals [Ruchty *et al*, 2010]; [Brandstaetter & Kleineidam, 2011]; [Galizia & Menzel, 1999], I sought to develop a calcium imaging platform to sample the activity of the majority of the antennal lobe simultaneously to reveal the logic of pheromone communication in ants. To this end, I adapted an imaging preparation previously used in drosophila, ants and honey bees [Galizia *et al*, 1997]; [Want *et al*, 2000] to directly visualize the activity of glomeruli within the antennal lobe. Nonetheless, given the negative result of not visualizing clear differences between different individuals, and the difficulty in obtaining clear responses from the presentation of different chemicals, we resort to alternative methodologies to improve odor delivery and loading of calcium indicators.

## A.4 The development of a calcium imaging platform.

*Identifying optimal Stimulus Condition by Antennogram Recordings.*

Electroantennography is a biological assay that records small voltage fluctuations in whole intact insect's antennas in response to the stimulation with volatile odorants. Although a theoretical explanation is still missing, the voltage changes between the tip and the base of the antenna are believed to be caused by the depolarization of olfactory receptor neurons in synchrony [Syntech 2004]. Electroantennogram recordings were used to identify defects in peripheral expression of genes in Drosophila Melanogaster mutants [Alcorta 1991], to identify two classes of odorant pathway (in the antenna and the maxillary pulp) [Ayer & Carlson, 1992] and to identify genes in involved in olfactory signal transduction [Ayer & Carlson, 1991]. In ants, it has been previously used to optimize low volatile compound delivery methods and to establish correlations between the caste and different strength responses to pheromones in different castes [Brandstaetter *et al*, 2010]; [Kleinedam *et al*, 2005].

We used electroantennogram recordings to validate two-photon microscope preparations by recording responses during hours to different chemicals and to isolate chemicals which elicit maximal response in the antenna. We hypothesize that highly responsive chemicals might be activating many different olfactory glomeruli and this, in turn, might facilitate validation of responses on the 2p preparation. Firstly, we validated the antennogram preparation by reproducing previous measurements on drosophila melanogaster. Next, we recorded depolarization in the antenna of pogonomyrmex barbatus and identified Acetone and Ethyl acetate as chemical producing high activation. These general activation smells (in combination with a random and broad palette of smells) were used to troubleshoot the two photon microscope preparation.

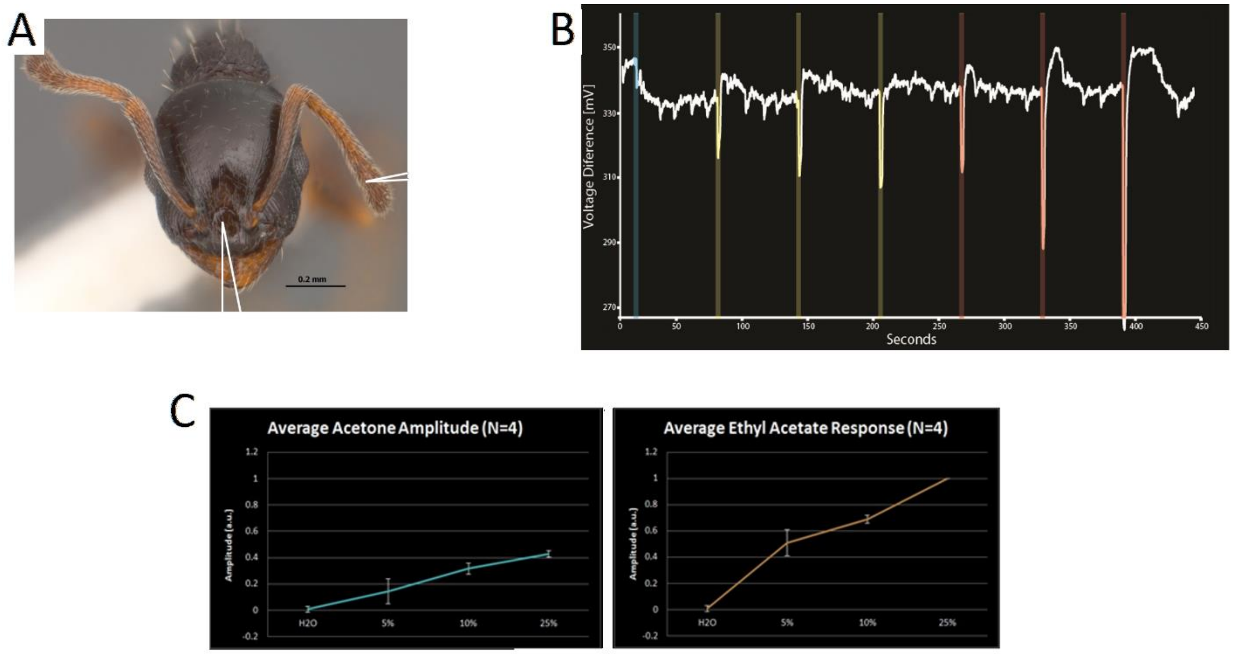


**Figure A1.6. Electroantenogram recordings.**

(a) Schematic representing the location at which recording electrodes contact the ant antenna. Ground electrode is inserted at the base of the antenna.
(b) Electroantennogram recordings depicting the presentation of two stimuli at different concentration. Yellow, acetone; red, Ethyl acetate.
(c) Average response to three different concentrations of stimuli in (b) after 5 trials.

*Improving Calcium Imaging Responses by delivering genetically encoded calcium indicators through viruses.*

Next, we sought to improve the signal to noise ratio in our two photon calcium imaging platform by delivering a genetically encoded calcium indicators (GECI) into the ant brain. A possible method to express exogenous genes into insect cells and embryos is viral delivery. Successful attempts have been made to deliver reporter genes (gfp) into drosophila S2 cells, silkworm embryos and adults and honeybee queens through the use of a baculoviruses or nucleopolyhedroviruses [Yamao *et al*, 1999]; [Kim *et al*, 2008]; [Ando *et al*, 2007]; [Ikeda *et al*, 2011].

Viruses that infect insect hosts provide ideal candidates to produce transgenic insects able to express genetically encoded calcium indicators. Candidate viruses belong to the family of Baculoviridae, and to the family of alpha viruses, like Sindbis virus. Both of these viruses are hosted by insects; silkworm in the case of Baculovirus; mosquitoes in the case of Sindbis. In addition, glycoprotein deleted rabies virus present an ideal option when their endogenous glycoprotein is replaced by the vesicular stomatitis viral glycoprotein, given the expanded host capabilities conferred by this enveloped proteins. In the case of Sindbis and Rabies virus, these two are rna-viruses, skipping the need of a promoter to drive their expression and high jacking the host cell machinery to translate their genes. Baculovirus is a DNA virus; in a typical expression system, exogenous genes are driven by an intrinsic viral promoter. However, we are not a priori certain about the functionality of this promoter in ant cells. For this reason, ant functional promoters were isolated. Describing the procedure by which these regulatory DNA sequences were identified is the goal of the next section.

Baculovirus was ultraconcentrated to achieve excessively high titer according to the manufacturer specification and subsequently purified through size exclusion chromatography [Transfiguracion *et al*, 2007]. In the same manner, rabies virus was pseudotyped and concentrated according to [Wickersham *et al*, 2010]; [Wickersham *et al*, 2013] and this method was used to concentrate Sindbis. Sindbis generation was performed according to [Foy *et al*, 2004] and [Lundstrom *et al*, 2012] (I am thankful to Thomas Reardon and Andy Murray from the Jessell laboratory, Kei Saotome and Sasha Sobolevsky from the Sobolevsky laboratory and Keneth Olson from Colorado State University for providing reagents and help in generating different viral variants). The examination of the proposed viruses is done in subsequent sections after we developed an ant cell culture assay.

*Promoter Region Identification*

A possible alternative for expressing exogenous genes in insect tissue is the identification of widely expressed promoter regions and the development of expression vectors. Efficient expression vectors have been used in the past to select transgenic embryos, to introduce foreign DNA into cells and to drive expression in electroporation experiments [Huynh & Zieler, 1999]. With the purpose of developing expression vectors useful for driving the expression of GECI in ants, we isolated different ant promoter regions from Pogonomyrmex Barbatus and tested them in cell culture.

We focused our efforts in isolating promoter regions for widely expressed genes such as Actin, Tubulin and a panneuronal marker termed Elav [O'Donnell *et al*, 1994]; [Natzle *et al*, 1984]; [O'Donnell & Wensick, 1994]; [Chung & Keller, 1990]; [Yao & White, 1994]. Using the existing genome assembly for Pogonomyrmex Barbatus, DNA regions spanning each gene was identified and the start codon mapped into the PBar genome. Based on the known promoter length for homologs genes in Dmel, candidate promoter regions were cloned and fused [Hobert 2002] to a reporter (gfp). Each DNA was introduced into two plasmid vectors, one necessary to construct a Baculovirus virus and the other one a common vector used to drive expression in Dmel insect cells.

Next, we tested such construct by transfecting them into S9 Schneider cells derived from drosophila embryos. Although ant cultured cells have not been developed, S9 cells provide a quick platform to validate our plasmid vectors for expression of the gfp reporter. Plasmids were introduced into S9 cells by using cellfectin according to the manufacturer instructions (minimal optimization of transfecting conditions was done using a positive control carrying the DMel actin

5c promoter). Tubulin and Actin resulted in the highest expression levels compared to Elav, although this result can only be stated qualitatively given the fact that no normalizing control was co-introduced with each plasmid. However, given its size (~1Kb), the tubulin promoter resulted ideal for expression system.

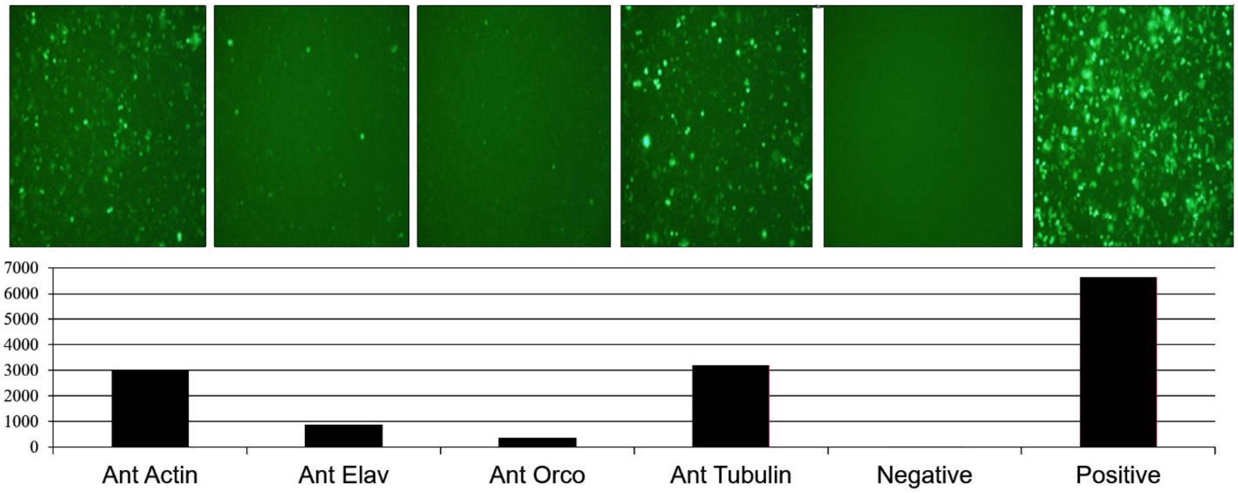


**Figure A1.7. Qualitative assessment of transfected constructs carrying ant promoters into S9 cells.**

Constructs carrying different ant promoters fused with eGFP at putative ATG positions. Transfection is performed using cellfectin reagent. Transfection conditions (not shown) were optimized using positive control (construct carrying Drosophila actin promoter expressing eGFP). It is worth noting that to perform a quantitative assessment of expression levels, a second channel should be include in which positive control would have been transfected in each condition. In this way, at every condition, a ratiometric measurement would have been performed.

(Top) Representative field of view (FOV) showing on green, gfp expressing cells.

(Bottom) Mean fluorescent of the FOV shown on Top (a.u.).

**Table A1.1. Primers used to clone ant promoter regions.**

| | |
|---|---|
| **Actin5c** | |
| Forward primer | ACCTTATTTGGTAGACAAGGTCG |
| Reverse primer | TACGAGCGCGGCAACTTC |
| Product length | 4327 |
| **Alpha Tubulin** | |
| Choice 1 | |
| Forward primer | TGGAGAGAAATGACTCGACCG |
| Reverse primer | GGCTTGTCCAACGTGGATTG |
| Product length | 1017 |
| **OrCo** | |
| Forward primer | CCATGCAGACGGCATAAACG |
| Reverse primer | ACACTTTATGCAAGTATTTGGACG |
| Product length | 3070 |
| **Elav** | |
| Forward primer | ATTTCCCCTTCTGTTCCGGG |
| Reverse primer | ACGACTGTGTCCATTCCGTT |
| Product length | 5387 |

*Ant Cell Culture Assay*

Cell culture systems are an excellent platform to study gene expression, widely used in molecular biology, genetic and biochemical studies [Bayne 1998]; [Barbara *et al*, 2008]; [Egger *et al*, 2013]; [Kreissl & Bicker, 1992]. We developed an ant cell culture with the purpose of testing the previously developed viruses and validate their efficacy in infecting ant cells. Given our limited availability of ant embryos, ant adult brain cells were dissociated according to protocol A1.2 and then plated. Different substrates (laminin, gelatin, glass and, plastic) were tested to guarantee optimal cell adhesion, with laminin resulting in better survival. Cell culture media was adapted from [Hunter, 2010], table A1.1 and guaranteed an average survival rate of 10 days.

**Protocol A1.2. Ant brain cell dissociation protocol.**

- Adult individuals from Pogonomyrmex colonies were used. For easiness, ants were glued to individual petridishes and around ~50 individuals are used each time.
- In a sterile, laminar flow hood, ants were surface sterilized with 70% ethanol.
- Samples were then rinsed three times with PBS.
- Sterile forceps were used to dissect the brain under cell medium.
- Brains were incubated in collagenase for 3 min. Next, collagenase was inactivated by addition of new medium (containing FBS).
- Next, brains were torn apart in medium by pipetting and then dispersed across multi-well plates, 24 wells (Costar®, Corning,NY) and incubates at 25°C temperature.
- After cells displayed attachment, usually within 1d, half media was exchanged at intervals of 2d.

**Table A1.1. Cell medium composition adapted from Honey bee, A. mellifera [Hunter, 2010]**

| | |
|---|---|
| Schneider’s Insect Medium | 150 ml |
| 0.06 ML-histidine solution (pH 6.35) | 200 ml |
| Fetal Bovine Serum (heat inactivated, 56°C for 30 min) | 50 ml |
| CMRL 1066 | 15 ml |
| Hanks’ Salts | 5 ml |
| Insect medium supplement (×10) | 2 ml |
| Gentamicin, units/1 µl/ml 1 µl/ml | |
| Total volume of medium | ~422 ml; |
| pH adjusted to 6.3–6.5 with 2 N HCl or NaOH | |

Finally, we tested different viruses in concentrations similar to the ones used to infect animals ($10^7$ virions) and observed that baculovirus and rabies viruses are able to infect ant cell cultures. Experiment performed on adult individuals result in negative expression.

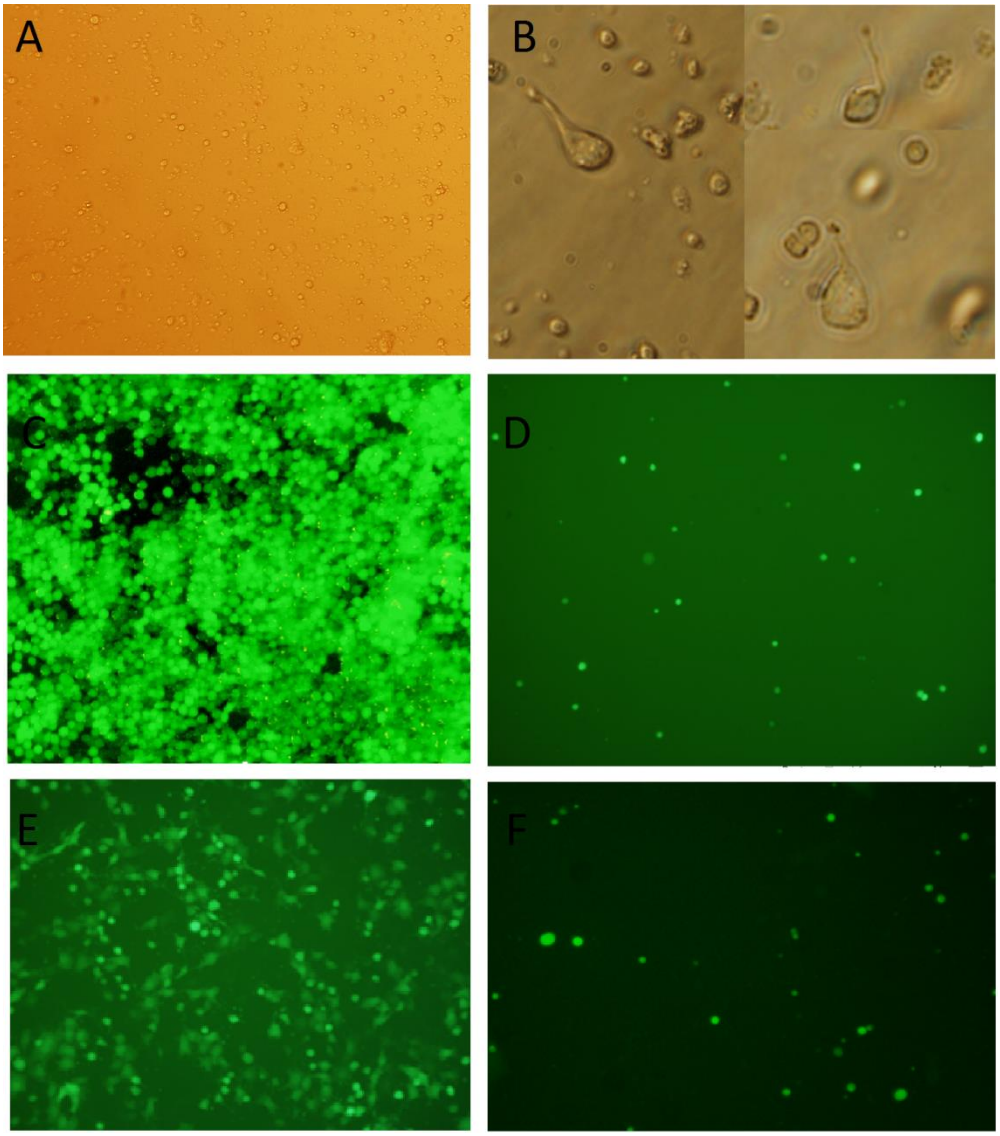


**Figure A1.8. Qualitative assessment of viral infection of ant brain cells.**

(a) Bright field depicting representative ant cell culture after 5 days.

(b) Zoom in FOV depicting ant cells growing processes.

(c) SF9 cells reporting in green infection with Baculovirus expressing GFP using the ant tubuling promoter.

(d) Same virus as in (c) infecting ant cells after 5 days. Only 10% of the cells express GFP.

(e) BHK cells reporting in green infection with pseudo-typed rabies virus carrying the VSV glycoprotein, expressing GFP.

(f) Same virus as in (e) infecting ant cells after 5 days. Only 10% of the cells express GFP.

# Appendix A - Bibliography

ADNS-3080 Avago Technologies. ADNS-3080 High-performance Optical Mouse Sensor.

Alcorta E. 1991. Characterization of the electroantenogramm in Drosophila melanogaster and its usefor identifying olfactory capture and transduction mutants. Journal of Neurophysiology 65, pp 702- 714.

Ando T., Fujiyuki T., Kawashima T., Morioka M., Kubo T., Fujiwara H. 2007. In vivo gene transfer into the honeybee using a nucleopolyhedrovirus vector. Biological and Biophysical Research Communications 352, pp 335-340.

Aso Y., Hattori D., Yu Y., Johnston R.M., Iyer N.A., Ngo T.T., Dionne H., Abbott L.F., Axel R., Tanimoto H., Rubin G.M.. 2014. The neuronal architecture of the mushroom body provides a logic for associative learning. Elife 23;3:e04577.

Ayer R.K., Carlson J. 1992. Olfactory physiology in the drosophila antenna and maxillary palp: acj6 distinguishes two classes of odorant pathways. Journal of Neurobiology 23, pp 965- 982.

Ayer R.K., Carlson J. 1991. Acj6: a gene affecting alfactory physiology and behavior in drosophila. PNAS 88, pp 5467-5471.

Bayne C.J. 1998. Methods in Cell Biology. Chp 10 Invertebrate Cell Culture considerations: Insects, Shellfish and worms. Academic Press.

Barbara G.S., Grunewald B., Paute S., Gauthier M., Raymond-Delpech V. 2008. Study of nicotinic acetylcholine receptors on cultured antennal lobe neurons from adult honeybee brains. Invertebrate Neuroscience 8, pp 19-29.

Brandstaetter A.S., Rossler W., Kleinedam C.J. 2010. Dummies versus air puffs: efficient stimulus delivery for low volatile odors. Chemical Senses 35, pp 323- 333.

Brandstaetter A.S., Kleineidam C.J. 2011. Journal of Neurophysiology 206, pp. 2437- 2449.

Billen J., Morgan D. 1998. Pheromone communication in social insects: Sources and secretions. Chapter 1. Westview Press.

Bonasio R., Zhang G., Ye C., Mutti N.S., Fang X., Qin N., Donahue G., Yang P., Li Q., Li C., Zhang P., Huang Z., Berger S.L., Reinberg D., Wang J., Liebig J. 2010. Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. Science 329, pp. 1068-1071.

Bransom K., Robie A.A., Bender J., Perona P., Dickinson M. 2007. High-throughput ethomics in large groups of Drosophila. Behavioral Ecology, pp 441- 447.

Blum M. 1969. Alarm Pheromones. Annual review of Entomology 14, pp 47-80.

Caron S.J., Ruta V., Abbott L.F., Axel R. 2013. Random convergence of olfactory inputs in the Drosophila mushroom body. Nature 497, pp. 113-117.

Chung Y.T. Keller E.B. 1990. Positive and negative regulatory elements mediating transcription from the drosophila melanogaster actin 5c distal promoter. Molecular Cellular Biology 10 (12), pp 6172-6180.

Datta S.R., Vasconcelos M.L., Ruta V., Luo S., Wong A., Demir E., Flores J., Balonze K., Dickson B.J., Axel R. 2008. The Drosophila pheromone cVA activates a sexually dimorphic neural circuit. Nature 27, pp 473-477.

Elsik C.G., Tayal A., Diesh C.M., Unni D.R., Emery M.L., Nguyen H.N., Hagen D.E. 2016. Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. Nucleic Acids Res. 44, pp. 793-800.

Egger B., van Giesen L., Moraru M., Sprecher S.G. 2013. In vitro imaging of primary neural cell culture from drosophila. Nature Protocols 8, pp 958-965.

Foy B.D., Myles K.M., Pierro D.J., Sanchez-Vargas I., Uhlirova M., Jindra M., Beaty B.J. and Olson K. 2004. Development of a new sindbis virus transducing system and its characterization in three culicine mosquitoes and two lepidopteran species. Insect Molecular Biology 13, pp 89-100.

Galizia C.G., Joerges J., Kuttner A., Faber T., Menzel R. 1997. A semi-in-vivo preparation for optical recording of the insect brain. Camponotus Rufipes. Journal of neuroscience methods 76, pp 61-69.

Galizia C.G., Sachse S., Rappert A., Menzel R. 1999. The glomerular code for odor representation is species specific in the honeybee Apis Mellifera. Nature 2, p 473-478.

Galizia C.G., Rossler W. 2010. Parallel Olfactory systems in insects: Anatomy and Function. The annual review of entomology 55, pp 399-420.

Greene M.J., Gordon D.M. 2003. Cuticular hydrocarbons inform task decisions. Nature 423, pp 32.

Guerrieri F.J., Nehring V., Jorgensen C.G., Nielsen J., Galizia C.G., d'Ettorre P. 2009. Ants recognize foes and not friends. Proceedings of the Royal Society B 282, pp 1-8.

Hobert O. 2002. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic C. elegans. Biotechniques 32, pp 728-730.

Holldobler B., Wilson E. O. 1990. The Ants. Belknap (Harvard University Press), Cambridge, MA.

Holldobler B., Morgan D.E., Oldham N.J., Liebig J. 2001. Recruitment pheromone in the harvester ant genus Pogonomyrmex. Journal of insect physiology 47, pp 369-374.

Holldobler B., Wilson E. O. 2008. The Superorganism. W. W. Norton, Publisher.

Honey Bee Genome. 2006. The Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee Apis mellifera.

Hunter W.B. 2010. Medium for development of bee cell cultures. In vitro Cellular Developmental Biology 46, pp 83-86.

Huynh C.Q. Zieler H. 1999. Construction of Modular and Versatile Plasmid vectors for the high-level expression of single or multiple genes in insects and insect cell lines. Journal Molecullar Biology 288, pp 13-20.

Ikeda T., Nakamura J., Furukawa S., Chantawannakul P., Sasaki M., Sasaki T. 2011. Transduction of baculovirus vectors to queen honeybees, Apis Mellifera. Apidologie 42, pp 461-471.

Jones, W.D., Cayirlioglu, P., Kadow, I.G., Vosshall, L.B. 2007. Two chemosensory receptors together mediate carbon dioxide detection in Drosophila. Nature 445, pp. 86-90.

Jenett A., Schindelin J., Heisenberg M. 2006. The virtual insect brain protocol: creating and comparing standardized neuroanatomy. BMC Bioinformatics 7, pp 1-12.

Jouni Sorvari, Pascal Theodora, Stefano Turillazzi, Harri Hakkarainen and Liselotte Sundström. 2007. Food resources, chemical signaling, and nest mate recognition in the ant Formica Aquilona. Behavioral Ecology, pp 441- 447.

Kirschner S., Kleinedam C.J., Zube C., Rybak J., Grunewald B., Rossler W. 2006. Dual Olfactory pathway in the Honeybee, Apis Mellifera. The journal of comparative neurology 499, pp 933-952.

Kim K.R., Kim Y.K., Cha H.J. 2008. Recombinant baculovirus-based multiple protein expression platform for Drosophila S2 cell culture. Journal of Biotechnology 133, pp 116-122.

Kleinedam C.J., Obermayer M., Halbich W., Rossler W. 2005. A macroglomerulus in the antennal lobe of Leaf-cutting ant workers and its possible functional significance. Chemical Senses 20, pp 383-392.

Kohatsu S., Koganezawa M., Yamamoto D. 2011. Female Contact Activates Male-Specific Interneurons that Trigger Stereotypic Courtship Behavior in Drosophila. Neuron 69, pp 498–508.

Kreissl S., Bicker G. 1992. Dissociated neurons of the pupal honeybee brain in cell culture. Journal of Neurocytology 21, 545-556.

Kuebler L.S., Kleinedam C.J. 2010. Distinct antennal lobe phenotypes in the leaf-cutting ant. The journal of comparative neurology 518, pp 352-365.

Legendre A., Miao X., Da Lage J., Wicker-Thomas C. 2008. Evolution of a desaturase involved in female pheromonal cuticular hydrocarbon biosynthesis and courtship behavior in Drosophila. Insect biochemistry and molecular biology 38, pp 244-255.

Li a., Liu T., Nie J., Guo L., Malicki J., Mara J., Holley SS.A., Xia W., Wong S.T.C. 2007. Detection of blob objects in microscopis zebrafish images based on gradient vector diffusion. Cytometry Part 71A, pp 835-845.

Lin H., Lai J.S., Chen Y., Chiang A. 2007. A map of olfactory representation in the Drosophila Mushroom Body. Cell 128, pp. 1205-1217.

Liu T., Li G., Nie J., Tarokh A., Zhou X., Guo L., Malicki J., Xia W., Wong S.T.C. 2008. An automated method for cell detection in Zebrafish. Neuroinformatics 5, pp 5-21.

Lundstrom K. 2012. Purification and concentration of alphavirus. Cold Spring Harbor Protocols.

Martin S.J., Vitikainen E., Helantera H., Drijfhout P. 2008. Chemical basis of nest-mate discrimination in the ant Formica Exsecta. Proc. of the Royal Soc. B 275, pp 1271-1278.

Martin S., Drijfhout F. 2009. A review of ant cuticular hydrocarbons. Journal of Chemical Ecology 35, pp 1151-1161.

McGurk D.J., Frost J., Eisenbraum E.J. 1966. Volatile compounds in ants: identification of 4-methyl-3-heptanone from pogonomyrmex ants. Journal of insect physiology 12, pp 1435-1441.

Montell C. 2009. A Taste of the Drosophila Gustatory Receptors. Cur. Opin. Neurobiol. 19, pp 345-353.

Morgan D.E. 2009. Trail pheromones of ants. Physiological Entomology 34, pp 1-17.

Mysore K., Subramanian K.A., Sarasij R.C., Suresh A., Shyamala B.V., VijayRahavan K., Rodruigues V. 2009. Caste and sex specific olfactory glomerular organization and brain structure in two sympatric ant species camponotus sericeus and camponotus compressus. Antropod structures and development 38, pp 485-497.

Nakagawa T., Sakurai T., Nishioka T., Touhara K. 2005. Insect sex-pheromone signals mediated by specific combinations of olfactory receptors. Science 307, pp 1638-1642.

Nakanishi A., Nishino H., Watanabe H., Yokohari F., Nishikawa M. 2010. Sex-specific antennal sensory system in the ant Camponotus japonicas: glomerular organizations of antennal lobes. Research in systems neuroscience 518, pp 2186-2201.

Natzle J.E., McCarthy B.J. 1984. Regulation of Drosophila \alpha – and \beta – Tubulin Genes during Development. Developmental Biology 104, pp 187 – 198.

Nishikawa M., Nishino H., Misaka Y., Kubota M., Tsuji E., Satoji Y., Yokohari F. 2008. Sexual dimorphism in the antennal lobe of the ant camponotus japonicus. Zoological Science, 25, pp 195-204

Optical Flow Website. http://www.bidouille.org/hack/mousecam.

Ruchty M., Helmchen F., Wehner R., Kleineidam C.J. 2010. Representation of thermal information in the antennal lobe of leaf-cutting ants. Front. Behav. Neurosci., Vol. 4, article 174.

Sakurai T., Nakagawa T., Mitsuno H., Mori H., Endo Y., Tanoue S., Yasukochi Y., Touhara K., Nishioka T. 2004. Identification and functional characterization of a sex pheromone receptor in the silkmoth bombyx mori. PNAS 101, pp 16653-16658.

Scott K., Brady R., Cravchik A., Morozov P., Rzhetsky A., Zuker C., Axel R. 2001. A chemosensory gene family encoding candidate gustatory and olfactory receptors in Drosophila. Cell 104, pp. 661-673.

Seelig J.D., Chiappe M.E., Lott G.K., Dutta A., Osborne J.E., Reiser M.B., Jayaraman V. 2010. Two-photon calcium imaging from head-fixed Drosophila during optomotor walking behavior. Nat. Methods. 7, pp 535-540.

Sharma K.R., Enzmann B.L., Schmidt Y., Moore D., Jones G.R., Parker J., Berger S.L., Reinberg D., Zwiebel L.J., Breit B., Liebig J., Ray A. 2015. Cuticular Hydrocarbon Pheromones for Social Behavior and Their Coding in the Ant Antenna. Cell Rep. 12, pp. 1261-1271.

Smith C.R., Smith C.D., Robertson H.M., Helmkampf M., Zimin A., Yandell M., Holt C., Hu H., Abouheif E., Benton R., *et al.* 2011. Draft genome of the red harvester ant Pogonomyrmex barbatus. Proc. Natl. Acad. Sci.108, pp. 5667-5672.

Syntech. 2004. Electroantennography, A practical introduction.

Suh, G.S.B., Wong, A.M., Hergarden, A.C., Wang, J.W., Simon, A.F., Benzer, S., Axel, R., Anderson, D.J. 2004. A single population of olfactory sensory neurons mediates an innate avoidance behavior in Drosophila. Nature 431, pp. 854-859.

Thistle R., Cameron P., Ghorayshi A., Dennison L., Scott K. 2012. Contact chemoreceptors mediate male-male repulsion and male-female attraction during Drosophila courtship. Cell 149, pp 1140-1151.

Tillman J.A., Seybold A.J., Jurenka R.A., Blomquist G.J. 1999. Insect pheromones -an overview of biosynthesis and endocrine regulation. Insect biochemistry and molecular biology 29, pp 481-514.

Transfiguracion J., Jorio H., Meghrous J., Jacob D., Kamen A. 2007. High yield purification of functional baculovirus vectors by size exclusion chromatography. Journal of virological methods 142, pp 21-28.

O’Donnell K.H., Chen C, Wensink P.C. 1994. Insulating DNA Directs Ubiquitous Transcription of the Drosophila Melanogaster \alpha 1 – tubulin Gene. Molecular and Cell Biology 14 (9), pp 6398 – 6408.

O’Donnell K.H., Wensink P.C. 1994. GAGA factor and TBF1 bind DNA elements that direct ubiquitous transcription of the \alpha 1-tubulin gene. Nucleic Acid Research 22 (22).

Ozaki M, Wada-Katsumata A, Fujikawa K, Iwasaki M, Yokohari F, Satoji Y, Nisimura T, Yamaoka R. 2005. Ant nestmate and non-nestmate discrimination by a chemosensory sensillum. Science 309, pp 311-314.

Peng H., Chung P., Long F., Jenett A., Seeds A.M., Myers E.W., Simpson J.H. 2010. BrainALigner: 3D registration atlases of Drosophila brains. Nature Methods 8, pp 493-500.

Vasquez G.M., Schal C., Silverman J. 2008. Cuticular hydrocarbons as queen adoption cues in the invasive argentine ant. The journal of experimental biology 211, pp 1249-1256.

Vosshall L.B., Wong A.M., Axel R. 2000. An olfactory sensory map in the fly brain. Cell 102, pp 147-159.

Vosshall L.B., Stocker R.F. 2007. Molecular Architecture of Smell and Taste in Drosophila. Annual Review of Neuroscience 30, pp. 505-533.

Wang L., Anderson D. 2010. Identification of an aggression-promoting pheromone and its receptor neurons in Drosophila. Nature 463, pp 227-232.

Wang J.W., Wong A.M., Flores J., Vosshall L.B., Axel R. 2003. Two-photon calcium imaging reveals an odor-evoked map of activity in the fly brain. Cell 112, pp 271-282.

Wang K., Gong J., Wang Q., Li H., Cheng Q., Liu Y., Zeng S., Wang Z. 2014. Parallel pathways convey olfactory information with opposite polarities in Drosophila. Proc. Natl. Acad. Sci. USA 111, pp 3164-3169.

Wickersham I., Sullivan H.A., Seung S.H. 2010. Production of glycoprotein-deleted rabies viruses for monosynaptic tracing and high-level gene expression in neurons. Nature Protocols 5, pp 595-606.

Wickersham I.R., Sullivan H.A., Seung H.S. 2013. Axonal and subcellular labelling using modified rabies viral vectors. Nature Communication 4, pp 2332-2340.

Wong AM, Wang JW, Axel R. 2002. Spatial representation of the glomerular map in the Drosophila protocerebrum. Cell 109, pp 229-241.

Yamagata N, Nishino H, Mizunami M. 2006. Pheromone-sensitive glomeruli in the primary olfactory centre of ants. Proceedings Biology Science 273, pp. 2219-2225.

Yamagata N, Nishino H, Mizunami M. 2007. Neural pathways for the processing of alarm pheromone in the ant brain. J Comp Neurol. 505, pp. 424-442.

Yamao M., Katayama N., Nakazawa H., Hayashi Y., Hara S., Kamei K., Mori H. 1999. Gene targeting in the silkworm by use of a baculovirus. Genes & Development 13, pp 511-516.

Yao K., White L. 1994. Neural Specificity of elav expression: defining a Drosophila promoter for directing expression to the nervous system. Journal of Neurochemistry 63, pp 41-51.

Zhou X., Slone J.D., Rokas A., Berger S.L., Liebig J., Ray A., Reinberg D., Zwiebel L.J. 2012. Phylogenetic and transcriptomic analysis of chemosensory receptors in a pair of divergent ant species reveals sex-specific signatures of odor coding. PLoS Genetics 8, e1002930.