Understanding Cellular Specialization Through Functional Genomics

Understanding Cellular Specialization through Functional Genomics

A dissertation presented by

Bradlee Nelms

to

The Committee on Higher Degrees in Biophysics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biophysics

Harvard University

Cambridge, Massachusetts

August 2015

Thesis advisor: Dr. Wayne Lencer Bradlee Nelms

Understanding Cellular Specialization through Functional Genomics

Abstract

The human body is composed of hundreds of specialized cell types, each fulfilling distinct functions that are together essential for normal tissue homeostasis. This thesis is aimed at identifying genes that contribute to to cell type-specific functions, with major projects focused on (1) a specialized epithelial transport pathway called transcytosis and (2) the challenge of measuring cell type-specific gene expression. In both projects, we applied high-throughput methods to narrow down from the ~25,000 protein coding genes to distinguish the subset that contribute to specialized cellular functions. Common themes include the development of enabling technology and the value of integrating diverse genomic datasets. The results described here implicate new genes in cell type-specific processes and provide a starting place for subsequent investigation into the individual genes and pathways.

In the first project, we performed an RNA interference (RNAi) screen to identify genes necessary for receptor-mediated transcytosis, a specialized endosomal pathway in epithelial cells. We developed high-throughput assays to measure the transcytosis of immunoglobulin G (IgG) across cultured epithelial cells in conjunction with gene knockdown. Then we selected a set of 582 candidate genes to screen using a combination of literature review and integrated high-throughput evidence, including

expression data, proteomics, and domain annotation. We knocked-down each of these candidates in parallel and identified many reagents that interfered with transcytosis. In small-scale validation assays, we confirmed a reproducible decrease in transcytosis after knocking down 7 genes with multiple independent reagents (7 confirmed out of 8 genes tested). The validated hits included genes with an established role in related pathways, such as *EXOC2* and *PARD6B,* and genes that have not been implicated in epithelial trafficking before, such as *LEPROT*, *VPS13C*, and *ARMT*.

In the second project, we developed an approach to identify genes expressed selectively in specific cell types, using a computational algorithm that searches thousands of microarrays for genes with a similar expression profile to known cell type-specific markers. Our method, *CellMapper*, is accurate without the need for cell isolation and can be applied to any cell type where at least one cell-specific marker gene is known. We demonstrated the approach for 30 diverse cell types, many of which have not been isolated for expression analysis in humans before. Furthermore, we explored the applicability of our method to infer causal relationships in genome-wide association studies (GWAS) and to investigate the transcriptional identity of a poorly understood cell type, *enteric glia*. We provided a user-friendly R implementation that will enable researchers from systems biology, molecular biology of disease, and population genetics to identify cellular localization of genes of interest or to expand the catalog of known marker genes for difficult-to-isolate cell types.

Table of Contents

Acknowledgements

I would first like to thank my advisor, Dr. Wayne Lencer, for his constant support and encouragement during my time in graduate school. Wayne has given me a remarkable amount of freedom to drive my own research projects, along with the guidance to insure that these projects were successful. I remember several years back when I approached Wayne with an idea to computationally predict which genes were broadly expressed in simple epithelial cells. Rather than persuade me against a project that was outside of his experience, he discussed the objectives and challenges with me and then put me in touch with two great collaborators – Drs. Curtis Huttenhower and Levi Waldron. With Wayne's guidance, this simple idea has since developed into an exciting paper, and left me with experience in bioinformatics and statistics that will be invaluable as I move into the next phase of academia. During graduate school, I have become fascinated with plant biology – due to the fundamental importance our green brethren have on human health and global sustainability – and I plan to move into plant research during my post-doctoral fellowship. Wayne has been an amazing mentor as I consider this next stage in my life: sharing the lessons he learned as his own diverse career path took him refugee camps in Cambodia to research labs in Boston, sending me to a national conference in plant biology to learn more about research in plants, and aiding in innumerable ways as I conducted my job search in a new field. I am leaving graduate school with a tough choice to make between several strong post-doc options, and I will be forever grateful to Wayne for his important role in getting me here.

I have also benefited in many ways from my scientific colleagues in the Lencer

laboratory and the greater Longwood medical area. Drs. Curtis Huttenhower and Levi Waldron – my collaborators on the *CellMapper* project – have been an irreplaceable resource as I entered the world of computational biology. I began graduate school as an experimental biologist and joined an experimental biology lab in an experimental biology division. When I started, I would have never dreamed that much of my thesis would be conducted on a computer. Curtis and Levi have guided me patiently as I – often clumsily – learned statistics and bioinformatics during the course of this project. Curtis has provided valuable advice on the overall direction of the project, and is now selflessly investing his time as a member of my defense committee. Levi has been my go to for all questions statistics, and has served as my primary mentor in computational research. From the Lencer laboratory: my benchmate Andrew Weflen for his frequent discussions about everything, from experimental design to data presentation to where to head to lunch if not Micheal's deli. Phi Loung for his good humor, clarity of thought, and terrifying work ethic. Lydia Kaoutzani for her support and friendship, and for introducing me to the necessity of coffee. Also my collaborator Dr. Meena Rao, who has been phenomenal to work with and has provided an excellent example of how to navigate an early stage career in academic research.

Among all the amazing resources in terms of equipment, scientific expertise, and other opportunities available at the Harvard Medical School, I have perhaps benefited most from my fellow graduate colleagues. Harvard – and the Biophysics program in particular – attracts an incredible mix of bright and engaged students. My graduate colleagues Max Staller, Rob Erdmann, and Geoff Fudenberg have frequently served as thought partners throughout every phase of graduate school, discussing challenges and

ideas ranging from the big picture to the nitty gritty. Max and Rob – thanks for the journal clubs, celebratory traditions, and consistent advice on all of my research. Geoff – thanks for the late night science chats and helping to make “the Franklin” home. I have also frequently found good collaborators in the Biophysics program, such as Luis Barrera and Chris McFarland. The Biophysics annual retreat offers an inspiring mix of biology – where in a span of a few hours you can hear talks on as diverse topics as rock respiring bacteria, p53 signaling in cancer, worm neuron opto-genetics, and quantum mechanical approaches to improve magnetic resonance imaging (MRI); after hearing about such broad and exciting research, I always return to lab invigorated and ready for the next phase of my own research. Thanks to Jim and Michele for putting together such an special program. And thanks to the Herchel Smith Graduate Fellowship Program for gathering a similarly diverse and engaged group of students for the periodic events and seminars.

My path in science began with valuable mentors from my undergraduate and early graduate education. It was Dr. M Thomas Record, my undergraduate advisor, who first introduced me to the obsession that is scientific research and inspired my path into graduate school. Dr. Christiane Wiese for her dedication to teaching and for having the courage to build a laboratory on undergraduates in an environment that does not favor such a move. Drs. Larabell and Parkinson for hosting me at the Berkeley National Laboratory for an engaging undergraduate internship. Thanks to my preliminary qualifying exam committee members, Drs. Fritz Roth, Gary Yellen, and John Dowling, and to my dissertation advisory committee and defense committee members Drs. David Clapham, Norbert Perrimon, Tom Kirchhausen, and Curtis Huttenhower – for taking the time to guide

me on projects unrelated to their own research and hosting many valuable discussions. And thanks to several outstanding professors during my first two years of graduate school, especially Drs. Stephen Harrison, Adam Cohen, and the “Computational and Functional Genomics” team – Drs. Martha Bulyk, Fritz Roth, and Shamil Sunyaev.

Finally, thanks to my friends and family for their encouragement and support during this process. This dissertation is dedicated to my grandfather, Warren Nelms, for his steadfast devotion to family and unwavering desire to serve others; I aspire to live my life in his example.

Preface

The Many Faces of Cellular Specialization

> *In each organ, the cells must have different characteristics, since such different substances are secreted within them. ... One can scarcely conceive that such an amazing diversity of products results from the activity of a single [structure] – the cell.*
>
> Rene-Joachim-Henri Dutrochet, 1824

Even as biologists were just beginning to recognize the *cell* as the fundamental structural unit in biology, many have been fascinated with cellular specialization. At the molecular level, cellular specialization is driven by differences in gene expression and activity: while nearly every cell in an individual contains the same genome, there is tremendous diversity in gene expression across cell types. The modern toolbox of functional genomics provides a powerful means to dissect cellular specialization because it allows thousands of genes to be interrogated at once, and thus facilitates the discovery of genes that contribute to

cell type-specific functions. This work is centered around two projects that apply functional genomics to identify genes important for cellular specialization.

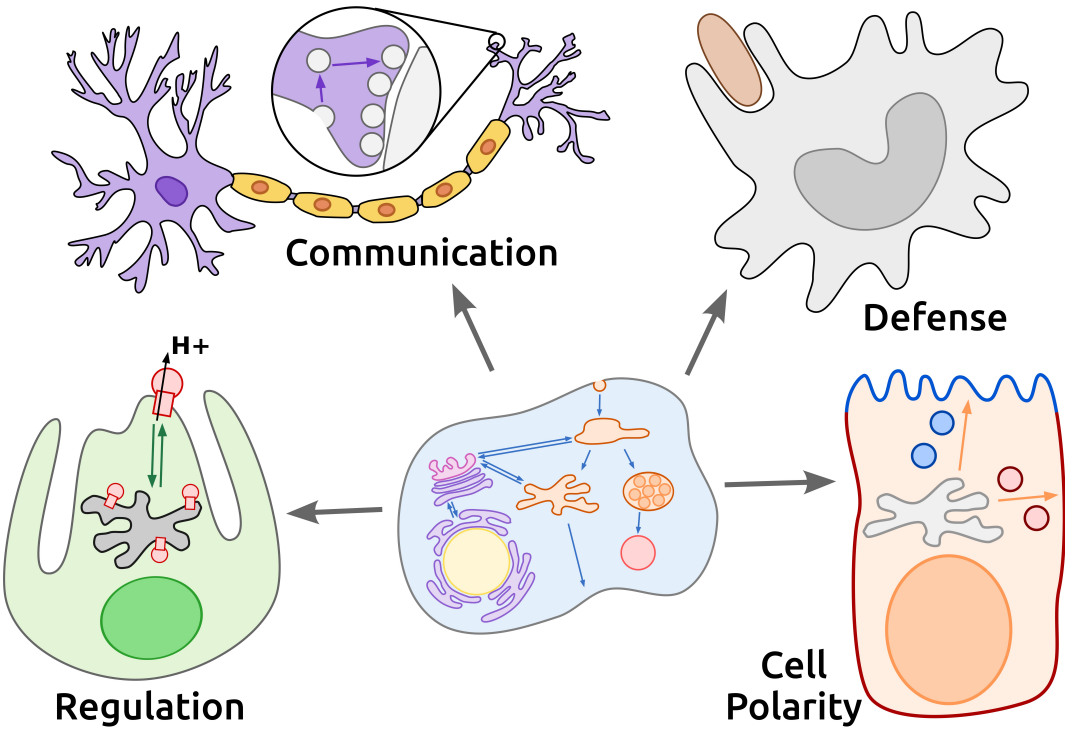


Figure 0.1: Examples of endosomal specialization in human cell types. Top left, synaptic transmission in neurons; top right, phagocytosis of invading bacteria by macrophages; bottom left, regulation of acid secretion in parietal cells lining the stomach; bottom right, maintenance and establishment of cell polarity in epithelial cells.

The first project focused on a specialized pathway called ***receptor-mediated transcytosis***, in which protein cargo is transported across epithelial cells through the endosomal system. The endosomes serve as a major hub for regulating the absorption and secretion of many compounds, as well as the precise control of plasma membrane composition. This organelle provides a fascinating case study in cellular specialization because it fulfills unique functions in many different cell types (**Fig 0.1**). In epithelial cells, the endosomes help to establish and maintain cell polarity, requiring additional endosomal

compartments and sorting mechanisms not found in other cell types[1]. How can the endosomes be reorganized to fulfill specialized functions such as transcytosis? The mechanisms of endosomal specialization in epithelia are not well understood, and one challenge has been that only a few genes that regulate epithelial-specific aspects of endosome trafficking are known. To help fill this gap, we conducted an RNA interference (RNAi) screen for genes that operate in receptor-mediated transcytosis. Our screen implicated several new genes in the process, expanding the catalog of genes that operate in this specialized epithelial transport pathway.

As one source of candidate genes for our RNAi screen, we hypothesized that many genes important for epithelial functions, such as polarized membrane trafficking, will be expressed selectively in epithelial cells. Examples of epithelia-specific trafficking proteins include *RAB25*, *RAB17*, and the clathrin adapter subunit *AP1M2* – all of which have demonstrated roles in polarized membrane transport[1]. However, there is currently no genome-scale database of genes expressed in many cell types, including epithelia. The second project focused on methods to estimate cell type-specific gene expression. We developed a new computational tool, called *CellMapper*, that leverages the wealth of available microarray data to predict which genes are expressed in specific cell types. Our approach was very successfully for epithelial cells, and we extended the analysis to many other cell types. CellMapper is effective without the need for cell isolation, and can be applied to any cell type where at least one cell-specific marker gene is known.

We then applied the CellMapper algorithm to other biological problems related to cellular specialization. In one application, we used the cell-specificity predictions of CellMapper to prioritize candidate genes in disease loci identified by genome-wide

association studies (GWAS). We identified several candidate disease genes that are selectively expressed in disease-relevant cell types, and therefore might contribute to cell type-specific functions that are disrupted during disease pathogenesis. In a second application, we applied CellMapper in parallel with RNA-Sequencing to identify genes selectively expressed in the cell type, *enteric glia*. We found that the expression profile of enteric glia is distinct from all other neural cell types, with some overlapping pathways but also many differences. These results suggest that enteric glia cannot be regarded as a close analog of any other cell type, and shed insight into the functions of this poorly understood cell type.

This document is organized in two sections, with each focused on one of the two major projects. In *Part I*, I describe the results of our RNAi screen for genes that function in receptor-mediated transcytosis: Chapter 1 provides background on transcytosis and endosomal trafficking and Chapter 2 describes the development and results of the reverse-genetic screen. Contributions include (i) the establishment of improved cell culture assays for transcytosis and (ii) the identification of genes that are necessary for receptor-mediated transcytosis. In *Part II*, I present the development and applications of CellMapper: Chapter 3 provides an overview of experimental and computational methods to estimate cell type-specific gene expression, Chapter 4 describes the CellMapper algorithm, and Chapter 5 discusses two biological applications of CellMapper. Contributions include (i) the creation of a new computational tool to predict genes expressed selectively in different cell types, and the application of this tool to (ii) prioritize candidate genes in human disease loci and (iii) reveal similarities and differences between enteric glia and other neural cell types.

An RNAi Screen for Factors that Regulate Membrane Transport in Polarized Epithelia

In the late 1970s, the fundamental question of how epithelial cells establish and maintain their polarized phenotype became experimentally approachable.

Rodriguez-Boulan, *et al.* 2005

concerning the impact of
the MDCK cell model
on epithelial cell biology

1

Endosomal Specialization in Polarized Epithelia

This chapter provides an introduction to vesicular trafficking in the endolysosomal system, with particular emphasis on endosome specialization in polarized epithelial cells. This background information is relevant to the RNAi screen described in Chapter 2.

1.0 Introduction

Epithelial cells are specialized to operate at the interface between vastly different environments. This is particularly true in mucosal tissues such as the intestine, where the epithelium separates a lumen saturated with microbes and microbial products from the underlying sterile tissue. Several features of epithelial cells contribute to the ability to form an effective barrier. For one, ***tight junctions*** span the border between individual cells, severely restricting the paracellular passage of microbes, food antigens, and other substances. Thus, passage across the healthy epithelium occurs primarily by transcellular routes that are tightly monitored and regulated[2]. Second, epithelial cells are polarized,

possessing separate ***apical*** and ***basolateral*** plasma membrane domains with distinct protein and lipid compositions. Cell polarity enables the epithelium to respond appropriately to signals from either surface and to transport molecules directionally across the epithelial layer[1].

For large molecules, for which no conducting channel exists, the primary route to cross the epithelium is transcellular vesicular transport, or ***transcytosis***[2]. Many roles for transcytosis have been documented *in vivo*. In addition to its physiological importance in the selective absorption or secretion of protein cargo, viruses[3] and bacterial toxins[4] co-opt the transcytotic pathway to traverse the epithelium and gain access to the underlying tissue. Transcytosis is also an essential pathway for epithelial cells to establish and maintain membrane polarity[1]. Despite the widespread importance of transcytosis in health and disease, the cellular machinery that directs membrane-protein traffic through the transcytotic pathway remains largely unknown.

1.1 FcRn-mediated transcytosis as a model system

The neonatal Fc receptor, ***FcRn***, binds to ***immunoglobulin G*** (IgG) and escorts it across the mucosal surfaces of the gut, lungs, and urogenital tract, where IgG operates in immune surveillance and host defense[5]. FcRn provides a valuable model for transcytosis because it physiologically transports IgG in both directions across the cell[6]. Notably, FcRn is the only established model of apical to basolateral (absorptive) transcytosis, a pathway poorly understood but significant for the uptake of antigens, microbial products, and protein therapeutics. In addition, many protocols and reagents are available to detect,

purify, and label IgG, facilitating the development of cell culture-based assays of FcRn-mediated IgG transport[6].

The extracellular domain of FcRn is structurally related to the Class I major histocompatibility complex (MHC). Like MHC class I, FcRn is an obligate heterodimer composed of a heavy chain and tightly associated β2-microglobulin (β2m). FcRn binds IgG in a pH dependent manner, with high affinity at low pH (< 6.5) but undetectable affinity at a pH greater than 7. This pH dependence allows FcRn to bind IgG within the acidified intracellular endosomes, and to release IgG at cell surfaces exposed to neutral pH[5]. FcRn continually cycles between the plasma membrane and endosomes, and is strictly sorted away from the lysosome[6].

In addition to its role in IgG transcytosis, FcRn also protects IgG from degradation in the blood stream. Fluid phase endocytosis, followed by lysosomal degradation, is a major source of protein turnover for soluble serum proteins. However, FcRn rescues internalized IgG from the lysosome by binding and recycling it back to the cell surface, thereby providing IgG with the one of the longest serum half lives of any protein[5]. The pharmaceutical industry has extensively studied the interaction between FcRn and IgG in an effort to extend the half life of therapeutic antibodies[7], leading to the identification of mutations in IgG that increase the affinity for FcRn[8,9]. In Chapter 2, we take advantage of one of these high affinity IgG mutants to develop more robust transcytosis assays.

1.2 Trafficking routes and sorting stations in polarized epithelia

The endosomes operate as a major sorting hub that connects the plasma membrane with

other membrane-bound intracellular organelles, such as the lysosome and trans-golgi network (TGN). After a protein is endocytosed from one surface, it can be recycled back to the same surface, transcytosed to the opposite surface, sent to the TGN, or directed to the lysosome for degradation. Protein cargo are routed to one of these possible destinations through a series of sorting stations[1] (**Fig 1.1A**). Newly internalized cargo first enters the early ***sorting endosomes***, where recycling and transcytotic cargo are sorted from degradative cargo[10] and directed to the recycling endosomes rather than the late endosome / lysosome. There is a distinct population of apical and basolateral sorting endosomes for cargo internalized from each cell surface[11], and the contents of these two endosomal populations do not mix[12].

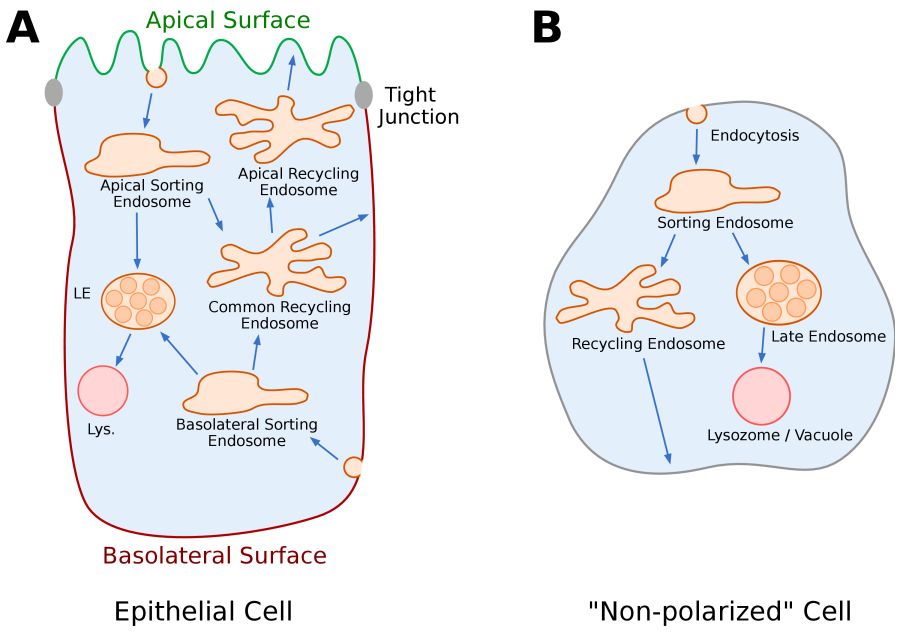


Figure 1.1: Endosome compartments and sorting stations. (A) Model of endosome compartments in polarized epithelial cells. (B) Model of endosome compartments in non-polarized cells.

Cargo that is to be returned to the cell surface moves next to the ***common recycling endosome*** (CRE). Endocytic pathways from both sides of the cell merge in the

CRE[13,14], and so this compartment serves as a key sorting hub where resident apical and basolateral membrane proteins are distinguished, and the recycling and transcytotic pathways diverge[14]. Cargo directed to the apical surface also makes an additional stop at the ***apical recycling endosome*** (ARE), which is either a separate compartment[15,16] or a distinct subdomain of the CRE[17]. The ARE is important for regulating access of membrane proteins to the apical surface[16], and can serve as a holding station for transmembrane proteins that are retained inside the cell until an appropriate environmental signal is received.

Several of the endosomal compartments described above are unique to polarized epithelia and not found in other cell types (compare **Fig 1.1A** and **Fig 1.1B**). These extra compartments are important for epithelial cells to accommodate the additional demands imposed by cell polarity, but very little is known about what distinguishes epithelia-specific endosomes from their analogs in other cell types. For instance, key regulators of sorting endosomes in non-polarized cells – such as *RAB5*, *EEA1*, and phosphatidylinositol 3-phosphate (PI3P) – are found on both apical and basolateral sorting endosomes and cannot differentiate these populations. A major problem in the field is that the cellular machinery that leads to epithelial-specific organization and regulation of the endosomes is largely unknown, and so it has been hard to dissect the pathway at a molecular level.

1.3 Molecular reactions in endosome trafficking

Numerous cellular factors control sorting and transport through the endosomes. Flux between endosomal compartments is primarily conducted by membrane-bound carriers

called ***vesicles***. The formation and subsequent delivery of these carrier vesicles to their correct destination can be divided into four stages[18]: budding, transport, docking, and fusion. In the first step, vesicle budding, a new vesicle is formed and pinched off from the membrane. Budding is mediated by a large complement of proteins, including ***coat proteins*** (e.g. clathrin) and others (e.g. caveolin, sorting nexins) that drive membrane curvature, ***adapter proteins*** that link specific cargo molecules to the coat complex (e.g. AP-2, Epsins), regulatory proteins that initiate and control the timing of budding reactions (e.g. ARF family GTPases), and proteins that catalyze vesicle scission (e.g. dynamin). After a new vesicle is formed, it is then transported to its destination by ***motor proteins*** (kinesins, unconventional myosins, dynein) that travel along the cytoskeletal filaments tubulin and actin. Next, the vesicle is recognized by ***tethering factors*** (e.g. *EEA1*, Exocyst complex) which capture the vesicle and direct it towards the destination membrane, helping to insure the specificity of vesicle targeting. Finally, the vesicle docks and fuses with the target membrane in a process catalyzed by the ***SNARE*** family proteins.

1.3.1 Control of endosome identity by Rabs and phosphoinositides

The timing and specificity of vesicular transport reactions are regulated by two major classes of molecule: the ***Rab family of small GTPases***[18] and ***phosphoinositides***[19]. Different members of these classes are distributed on specific endosomal compartments, serving as the primary molecular determinants of compartment identity. For instance, the early sorting endosome is populated by *RAB5* and phosphoinositol 3-phosphate (PI3P). *RAB5* and PI3P together recruit and activate other proteins involved in vesicular transport,

such as tethering and coat proteins, and thereby define the set of molecules that are present and active on the sorting endosome.

The localization of Rabs and phosphoinositides is controlled, in turn, by accessory proteins. Rabs exist in two distinct forms, an active form that is bound to GTP and an inactive form that is bound to GDP. Rab proteins are converted between the active and inactive states by proteins called guanine nucleotide exchange factors (**GEFs**), which activate the Rab by exchanging GDP for GTP, and GTPase-activating proteins (**GAPs**), which inactivate the Rab by stimulating the hydrolysis of GTP. Rab proteins often recruit their own GEFs, creating a positive feedback loop that allows a burgeoning endosomal compartment to be rapidly populated by a specific Rab isoform[18]. Rab proteins can also recruit GAPs and inactivate GEFs for other Rab isoforms, breaking a preexisting positive feedback loop and allowing one endosomal compartment to mature into another – a process called Rab conversion[20]. The situation is similar for phosphoinositides, as a series of kinases and phosphatases are recruited to convert one phosphoinositide (e.g. PI3P) to a different phosphoinositide (e.g. $PI(3,5)P_2$), changing the molecular identity of the membrane[19]. Thus, a network of positive and negative feedback loops insures the stability of mutually exclusive populations of Rabs and phosphoinosites, and therefore distinct endosomal compartments.

Rabs and phosphoinositides are important for establishing the specialized endosomal compartments and plasma membrane domains in polarized epithelia. Phosphoinositides are asymmetrically distributed between the apical and basolateral surface domains, with $PI(4,5)P_2$ enriched on the apical surface[21] and $PI(3,4,5)P_3$ on the basolateral surface[22]. The asymmetric distribution of phosphoinositides is functionally

important for cell polarity, as many proteins become mis-localized when the distribution of these phosphoinositides is altered[21,22]. Several Rabs are located on specific endosomal compartments; *RAB8*, *RAB11*, *RAB17*, and *RAB25* are all found on the recycling endosomes[1,23], with several enriched on the ARE relative to the CRE[15]. Chronic knock-down of these Rabs disrupts the trafficking of several transmembrane receptors[1]. For FcRn, transcytosis is inhibited by knock-down of *RAB25* and basolateral recycling by knock-down of *RAB11*[6].

1.3.2 Sorting signals and receptors in polarized epithelia

Epithelial cells continually sort apical and basolateral proteins from each other within the secretory and endocytic pathways. Clathrin is important for sorting many single-spanning transmembrane proteins to the basolateral surface. Indeed, some of the first basolateral sorting signals to be discovered resemble the YXXΦ motif known to bind to the ***adaptin*** (AP) family of clathrin adapters[1,24]. Chronic knock-down of the clathrin heavy chain resulted in the mislocalization of many basolateral proteins to the apical surface, while apical proteins were unaffected[25]. The function of clathrin in basolateral polarity is mediated, in part, by the adaptin subunit *AP1M2*. Adaptin mu subunits, such as *AP1M2*, bind to YXXΦ motifs in the cytosolic tail of membrane proteins and link these proteins to the clathrin coat. Several lines of evidence support a central role for *AP1M2* in basolateral sorting in epithelial cells: it is expressed specifically in epithelia, resides on recycling endosomes[26], and is necessary for the basolateral polarity of transmembrane receptors such as *LDLR* and *TFRC*[27]. However, much remains to be discovered. The factors that

recruit *AP1M2* to recycling endosomes are not known. Furthermore, additional sorting pathways must exist for proteins that do not depend on clathrin for basolateral targeting.

Sorting mechanisms for apical proteins are less well defined. Sorting signals have been been identified in the cytosolic, transmembrane, and extracellular domains of apically localized proteins[28], but very few molecular factors that recognize these signals are known. Some glycosylated apical proteins may be sorted by specific receptors, such as the lectin *LGALS3*[29]. However, the identification of sorting receptors for many apical proteins has been illusive. One idea is that these proteins are sorted by incorporation into microdomains called ***lipid rafts***[30]. The apical membrane is enriched for components – such as cholesterol, glycosphingolipid, and glycosylphosphatidyl inositol (GPI) anchored proteins – which segregate into macroscopically visible domains in model lipid bilayers and co-purify in a detergent resistant fraction of cell extracts[31]. The lipid raft hypothesis proposes that these lipids and proteins also cluster together in cells and form microscopic domains called 'rafts', which can then be incorporated together into apically bound carrier vesicles.

Throughout this introduction, I have pointed out several gaps in our knowledge about endosome trafficking in polarized epithelia. In the current era of biology, there is significant power in knowing which genes operate in a pathway. Many tools are available to alter the activity of a gene once its identity has been determined. Thus, if genes that act in multiple stages of a pathway are identified, it is possible to interfere with each in turn and dissect how the pathway fits together. The overarching goal of my first graduate project has been to uncover genes that regulate protein transport across epithelial cells, providing a foothold with which to dissect pathway structure.

2

Identifying Genes that Direct IgG Transport Across Polarized Epithelial Cells

This chapter describes a cell based RNA interference (RNAi) screen to identify genes that act in receptor-mediated transcytosis. I conceived and designed this project in collaboration with my advisor Wayne Lencer, and carried out all experiments either individually or with the support of Natasha Furtado Dalomba, a talented undergraduate student, and Sean Johnston, a robotics specialist at the Institute of Chemistry and Cell Biology, Longwood (ICCB-L).

2.0 Introduction

Large (macromolecular) cargo, such as immunoglobulins and chaperone-dependent nutrients, cross epithelial barriers in a receptor-mediated process known as transcytosis[2]. A major problem in the field is that much of the cellular machinery that orchestrates transcytosis is not known, and so it has been hard to dissect the pathway at a molecular level. To fill this gap, we conducted a targeted RNAi screen for genes that operate in

FcRn-mediated IgG transcytosis. We identified many cases where the depletion of specific genes caused a reproducible decrease in IgG transport, including seven subunits of the exocyst complex, the polarity protein *PARD6B*, and several genes that have not been linked to membrane trafficking in polarized epithelia before, such as *LEPROT* and *VPS13C*.

2.1 Developing an assay of receptor-mediated transcytosis compatible with high-throughput screening (HTS)

The Lencer laboratory has previously established biochemical assays to measure IgG transport using the Madin-Darby Canine Kidney (MDCK) epithelial cell line. In this experimental system, MDCK cells stably co-express human β2m and the FcRn heavy chain tagged with the hemaglutinin (HA) epitope at the N-terminus and EGFP at the C-terminus. This FcRn fusion construct is correctly localized to the early endosomes and can functionally bind IgG and transport it across the cell[6]. To measure transcytosis, cells are grown on semipermeable support filters (**Fig 2.1**), which provide independent experimental access to both the apical and basolateral sides of the monolayer. Then IgG is added to one side of the monolayer at pH 6.0 (to facilitate FcRn-dependent uptake), and the amount of IgG to pass through the filter after a period of time is quantified by an enzyme-linked immunosorbent assay (ELISA). In preparation for an RNAi screen, several modifications were made to the original MDCK model.

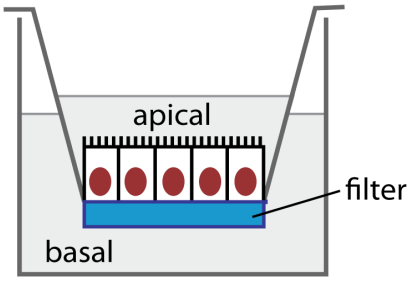


Figure 2.1: Schematic of the transwell experimental system for epithelial transport assays.

2.1.1 Increasing assay sensitivity with Fc fusion proteins

One challenge of the original transcytosis protocol was that the ELISA used to measure IgG concentration required a large amount of input material, and was not sensitive enough to accurately measure concentration with the small sample volumes of a 96-well plate. To

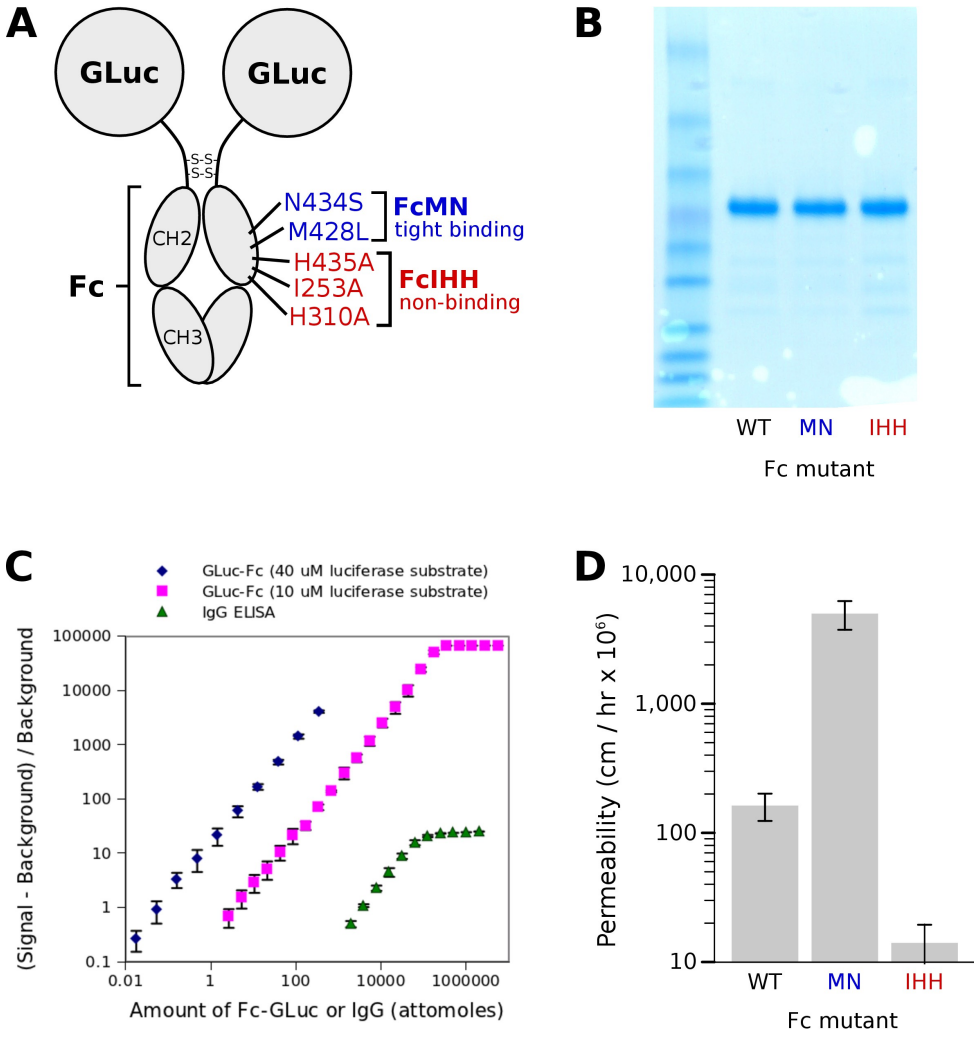


Figure 2.2: Increasing assay sensitivity with Fc fusion proteins. (A) Schematic of the hGLuc-Fc fusion protein and the location of mutations that increase (FcMN) or abolish (FcIHH) binding affinity for FcRn. (B) Purity of fusion proteins produced in CHO cells. (C) Log-log titration curves showing the dynamic range of GLuc-Fc compared to the old IgG ELISA; sensitivity of the GLuc assay can be increased by raising the luciferase substrate concentration. (D) Functional validation of the hGluc-Fc fusion proteins in FcRn transport assays.

overcome this, we constructed a recombinant protein in which the Fc (FcRn-binding) domain of IgG1 is linked to a humanized version of Gaussia luciferase (hGLuc; **Fig 2.2A**). GLuc is the brightest known luciferase, and provides a convenient enzymatic reporter for our transcytosis studies. The hGLuc-Fc fusions were produced in CHO cells, resulting in a single strong band on an SDS-PAGE gel (**Fig 2.2B**). The new hGLuc-Fc fusion proteins can be detected with 5-log greater sensitivity and at least a 4-fold greater dynamic range than previous assays for full length IgG (**Fig 2.2C**). The hGluc-Fc fusion also simplifies the assay because its concentration can be measured directly, alleviating the need for the many binding and washing steps of an ELISA. These advances made it possible to measure protein concentration in a 96-well format.

In addition to improving the ease of detecting FcRn-dependent cargo, we took steps to increase the biological signal-to-noise of our transport assays. As mentioned in Chapter 1, FcRn is responsible for the long serum half-life of IgG, and a significant amount of work has gone into modulating FcRn-binding affinity to increase the half-life of therapeutic antibodies. Several Fc mutants have been identified that bind to FcRn with increased affinity at low pH, yet still release effectively at neutral pH. We reasoned that such a mutant might exhibit higher FcRn-mediated transport *in vitro*, boosting the signal-to-noise of our assays. To test this, we introduced an M428L/N434S double point mutation in the FcRn binding site of our fusion proteins. This mutation has previously been shown to increase FcRn-binding affinity *in vitro* and boost FcRn-mediated transport *in vivo*[9]. Indeed, the new mutant – which we designate hGLuc-FcMN – is transported across MDCK monolayers 20 times more efficiently than wild type hGLuc-Fc (**Fig. 2.2D**). For comparison, an hGLuc-Fc fusion that contains inactivating mutations in the FcRn-binding

site (I253A/H310A/H435A; designated hGLuc-FcIHH) is not transported efficiently across MDCK monolayers (**Fig. 2.2D**, left). The FcIHH mutation abolishes binding affinity to FcRn but not to other immune Fc receptors[6], and thus provides a clean measure of FcRn-independent permeability.

2.1.2 Protocols for robust gene knock down in MDCK cells

To prepare our MDCK model for a cell based RNAi screen, another challenge was the need to develop protocols for consistent and efficient gene knock-down using the MDCK cell line. We chose to use endoribonuclease-prepared siRNAs (esiRNAs) because they are inexpensive to prepare and have been shown to produce robust knock down in mammalian cell culture[32]. EsiRNAs are prepared enzymatically as follows (**Fig. 2.3A**): first, primers are designed to amplify a 400-600 bp region of a target gene from a cDNA library, appending T7 promoter sites to each end. Second, the PCR product is transcribed in vitro to produce long double stranded RNA (dsRNA). Finally, the dsRNA is spliced into 21 bp fragments by a bacterial RNase, creating a heterogeneous pool of siRNAs. The resulting esiRNA reagent contains a diverse pool of sequences targeting the same gene, which has been experimentally shown to produce more consistent knock down and less off target effects than single, chemically synthesized siRNAs[33].

To test the esiRNA strategy, we prepared esiRNAs targeting *Gaussia* luciferase (as a negative control) and FcRn. The corresponding DNA fragments were amplified by PCR, and esiRNAs were synthesized (**Fig. 2.3B**). We then optimized conditions for gene knock down in MDCKs using the lipid based transfection reagent, lipofectamine RNAiMAX. With

our optimized transfection protocol, FcRn-GFP expression was reduced by ~85% with esiRNAs targeting the receptor (**Fig 2.3C**, right), and this knock down produced a corresponding functional decrease in IgG transport (**Fig 2.3C**, left). This finding has been reproduced using independent preps of each esiRNA, demonstrating that both the synthesis and knock down are robust.

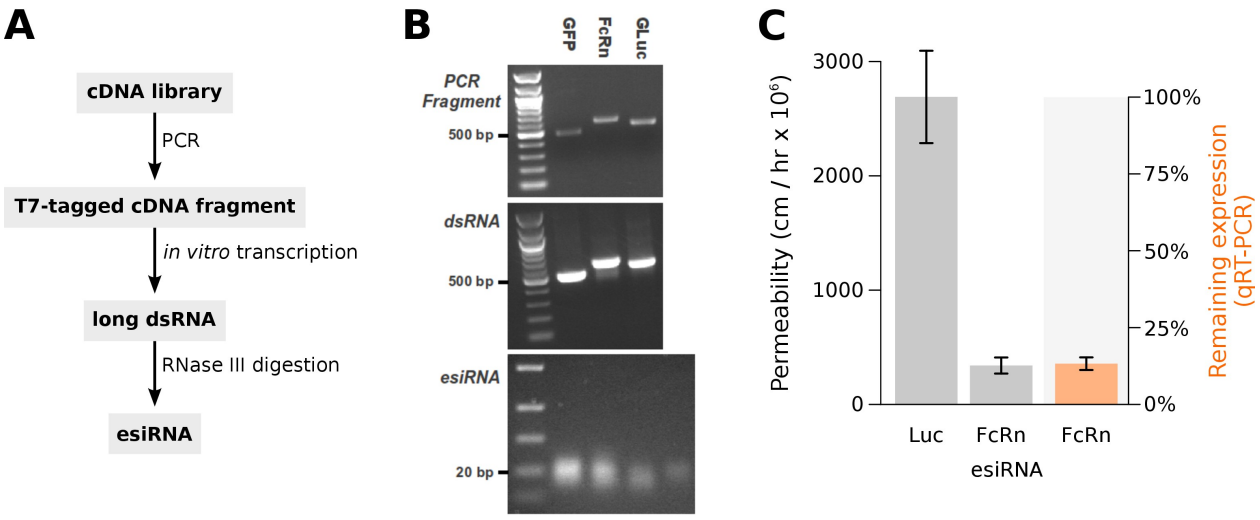


Figure 2.3: Preparation and validation of esiRNAs for gene knock down in MDCK cells. (A) Flow chart of the esiRNA synthesis protocol. (B) Products from each stage of synthesis were run on agarose gels; a chemically synthesized 21 bp siRNA standard is shown in lane 4 of the bottom gel. (C) Functional knock down of FcRn using esiRNAs. Left, apical to basolateral IgG transport assay after transfection with a non-targeting esiRNA (Luc) or an esiRNA targeting FcRn. Right, qRT-PCR shows that FcRn was knocked down by ~85% with the targeting esiRNA.

2.2 esiRNA library design and construction

2.2.1 Selecting candidate genes to target

To select candidate genes for our screen, we adopted a strategy designed to take advantage of known biology without sacrificing the potential to discover novel genes

and/or gene families. First, we combed through several major reviews on membrane trafficking in epithelia[1,23,31,34–39], and compiled a list of 143 potentially relevant genes mentioned by these reviews (the “Literature Curated Geneset”). Second, we searched for genes implicated in epithelial endosomal trafficking by several high-throughput sources of evidence:

- *Location*: the gene, or an orthologous yeast gene, has been localized to the endosomal system by proteomics[40–42] or GFP-tagging[43]
- *Domain*: the gene contains at least one protein domain (eg Rab GTPase, FYVE-finger, BAR) related to membrane trafficking or targeting to endosomes
- *Expression*: the gene was classified as epithelia-expressed by our cell type deconvolution algorithm *CellMapper* (described in Chapter 4 of this document)
- *Phenotype*: the gene was identified in either of two genome-wide RNAi screens for regulators of endocytic trafficking[44,45]

Each of these catagories of evidence highlighted a significant number of established epithelial trafficking genes (**Figure 2.4A**). Of the categories, “Location” was the most enriched for literature curated epithelial trafficking genes (12.3 fold more literature curated genes than expected by chance), and “Domain” recovered the highest percentage of literature curated genes (91.6%). We considered any gene highlighted by ≥2 categories of evidence to be potential candidates for our screen. This multi-evidence strategy resulted in a set of candidates that was 13.9 fold enriched in literature curated genes – more than any single evidence category – and included 89 out of 143 literature curated genes (62% coverage), second only to domain annotation. The strategy also highlighted many genes

with unknown function, and so the high-throughput approach identified candidates that would not have been predicted by literature review alone.

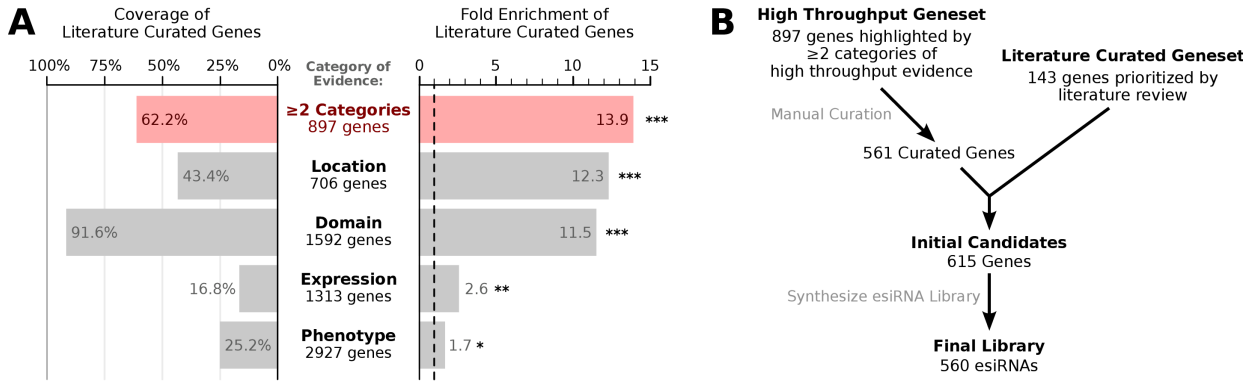


Figure 2.4: Selecting candidate genes for an RNAi screen. (A) Evaluating how effectively different high-throughput categories of evidence recover a literature curated set of 143 epithelial trafficking genes, showing both % coverage and fold enrichment. (B) Schematic of candidate selection and esiRNA synthesis.

We then manually curated these candidates, removing any gene if the high-throughput evidence could be fully explained by the gene's known function, and this function was unrelated to regulating membrane transport. We also included literature-curated genes that were missed by the high-throughput evidence, resulting in a combined list of 637 candidate genes, 22 of which were selected for a pilot screen and the remaining 615 for the full screen (**Fig 2.4B**).

2.2.2 esiRNA library construction

The starting point for esiRNA synthesis is a complementary DNA (cDNA) library. We constructed a cDNA library from polyA+ RNA purified from our MDCK cell line using the 'template-switching' cDNA synthesis protocol described by Pinto and Lindblad[46], with the

addition of Trehalose and Betaine to facilitate amplification of long transcripts[47]. This cDNA synthesis protocol resulted in substantial library diversity, with ~2 x 10^9 total cDNA molecules as estimated by dilution PCR and 75% of clones representing full length transcripts. The cDNA library was then amplified by emulsion PCR[48] and provided to Eupheria Biotech in Germany. Eupheria specializes in esiRNA synthesis, and produced esiRNAs for all genes we chose to target in this screen. In total, esiRNA synthesis was successful for 22 out of 22 genes selected for the pilot screen (100%) and 560 out of 615 genes selected for the full screen (91%).

2.3 RNAi screen

Our high throughput assays were conducted at the Institute of Chemistry and Chemical Biology - Longwood (ICCB-L) screening facility. Both esiRNA transfection and transport assay were performed with the aid of robotics, allowing for tight control of the timing of each step.

2.3.1 Pilot screen

To validate the high throughput transcytosis assay, we performed a pilot screen against 24 genes. The pilot included a non-targeting control (Luciferase), off-target controls known to affect general cell health (*KIF11*, *COPB1*, *STX5*), and negative controls that are located on endosomes but are not expected to directly regulate transcytosis (*TFRC*, *MYD88*). In addition, 2 distinct esiRNAs were synthesized against five genes in order to test the reproducible different reagents targeting the same gene.

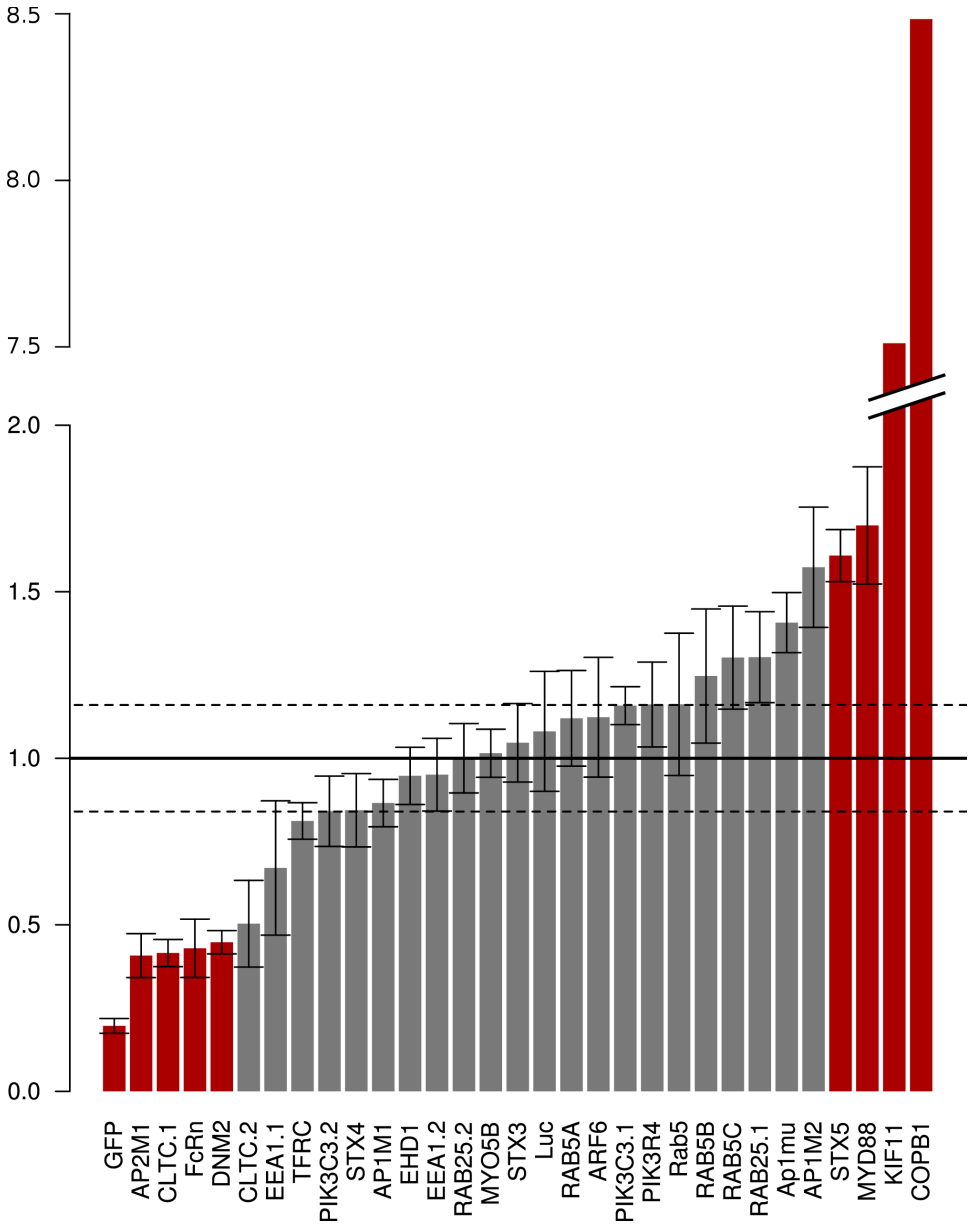


Figure 2.5: Results from the pilot screen, showing basolateral to apical transport assays. The horizontal solid and dotted lines are mean and standard deviation of the negative control wells. Red indicates a statistically significant difference compared to negative controls (two-tailed t-test adjusted for multiple hypothesese by holm's method). However, there was noticeable systematic within-plate variability (edge effects), and so statistical differences should be interpreted with caution.

Results from the pilot are shown in **Figure 2.5**. The two positive control esiRNAs targeting FcRn-GFP (GFP, FcRn) produced a significant decrease in Gluc-Fc permeability, demonstrating our ability to inhibit the FcRn-dependent transport pathway. We also observed a clear decrease in permeability when targeting genes required for clathrin-mediated endocytosis (*CLTC*, *AP2M1*, *DNM2*), in conjunction with a redistribution of FcRn to the cell surface (**Fig 2.6A**). On the other hand, esiRNAs against KIF11 and COPB1 produced a substantial increase in permeability. Both KIF11 and COPB1 are active in essential cellular processes (mitosis and ER to golgi transport, respectively), and knocking down these genes in our MDCK model severely disrupts cell health, preventing the cells from establishing a confluent monolayer (**Fig 2.6B**).

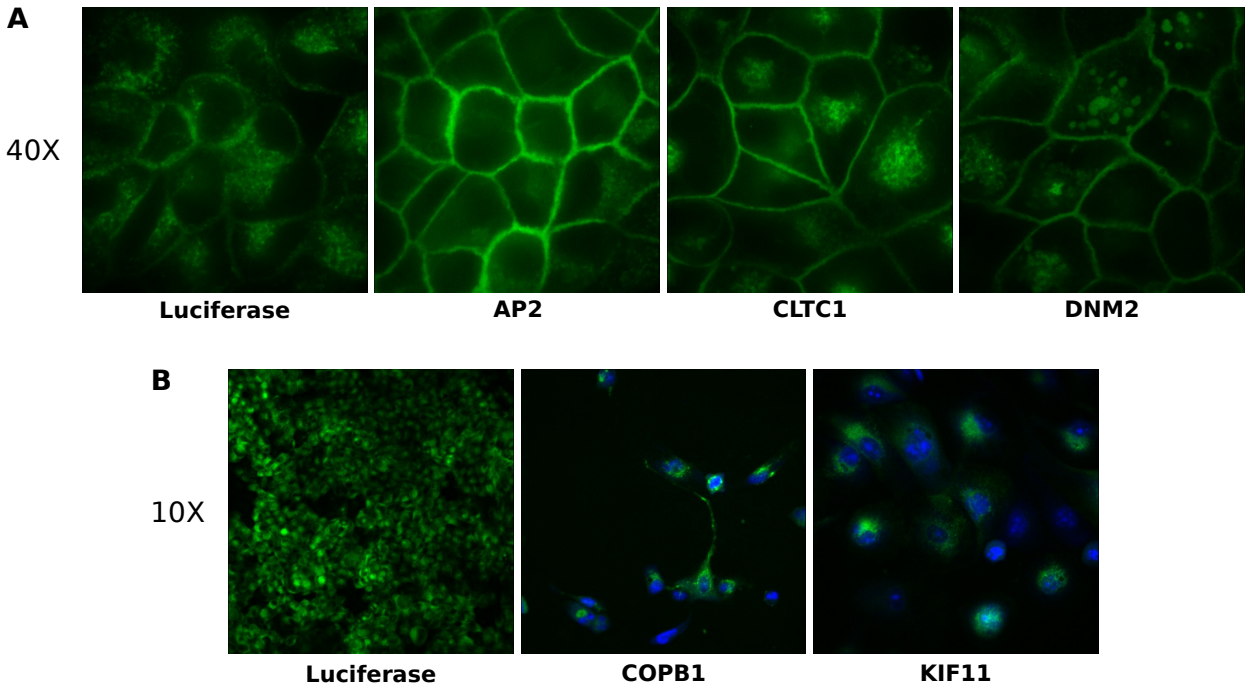


Figure 2.6: Microscopy on fixed wells from the pilot screen. (A) 40X view of FcRn-GFP in wells treated with an esiRNA against Luciferase (non-targeting), or esiRNAs against components of the clathrin-mediated endocytosis machinery. (B) 10X view of FcRn-GFP in wells treated with an esiRNA against Luciferase, or esiRNAs that strongly disrupted cell health.

To assess knock down efficiency under screening conditions, we lysed several wells from the pilot screen and measured gene expression by qRT-PCR (**Fig 2.7**). For this analysis, we selected the five genes targeted by multiple esiRNAs, comparing expression in wells treated with each targeting esiRNA against those treated with the non-targeting esiGLuc control. Knock down ranged from 60-90%, with every gene reduced ≥75% by at least one esiRNA. There was a general tendency for the first targeting esiRNA to have a larger effect than the second; this can be explained because the region targeted by the esiRNAs was optimized by a computational algorithm, and so the first reagent against each gene has been selected for the optimal nucleotide sequence for gene silencing.

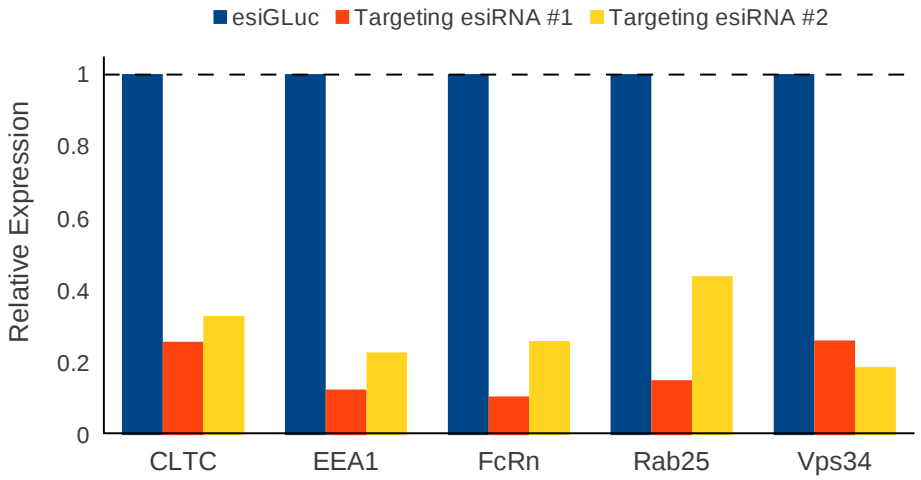


Figure 2.7: Quantitative reverse transcriptase PCR (qRT-PCR) of 5 genes in MDCKs treated with a non-targeting esiRNA (esiGLuc) or two different targeting esiRNAs. EsiRNAs against Gluc and FcRn were synthesized in the Lencer lab, all other esiRNAs were synthesized by Eupheria.

In the pilot screen, we normalized each plate to negative control wells. However, we found that the 6 negative control wells included in the pilot were not enough for accurate plate-wise normalization. We also observed a tendency for systematic variability, such as edge effects, within a single plate. Therefore, in order to adjust for systematic

effects within and between plates, we increased the number of control wells in the full screen, as discussed below.

2.3.2 Full screen

The entire esiRNA library was screened over a period of three sessions at ICCB-L. Similar to what was observed in the pilot screen, the raw data displayed substantial systematic variability such as batch and edge effects. However, several aspects of our experimental design made it possible to control for this systematic variability and identify reproducible hits. First, we included a large number of control wells distributed across each plate (**Fig 2.8A**); this layout was designed to fit statistician's recommendations for high-throughput screens[49]. Second, we performed the screen with substantial replication (**Fig 2.8B**): every library plate was screened in triplicate in both transport directions, and protein concentration was measured in duplicate for every plate.

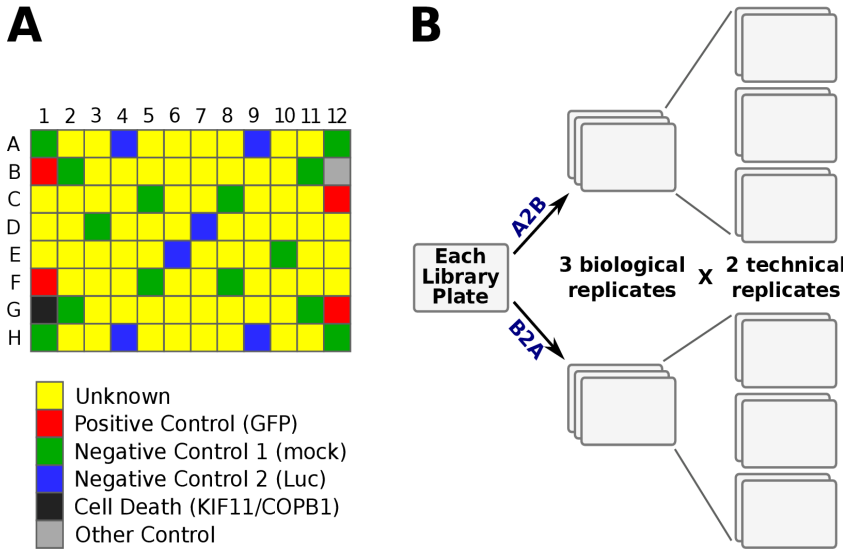


Figure 2.8: Plate layout and replication design. (A) Plate layout showing the position of all controls. (B) Diagram of replication schema.

The raw data were normalized using a linear mixed-effects model[50], resulting in a substantial decrease in variability. Normalized biological replicates displayed a high degree of reproducibility, with an R^2 of 0.907 for basolateral to apical transport and 0.966 for apical to basolateral transport (**Fig 2.9**). Replicate values for each esiRNA were then pooled together and converted into a Z-score, which reflects the number of standard deviations of the phenotype relative to the mock transfection controls. More details about the normalization and summarization can be found in Appendix A.

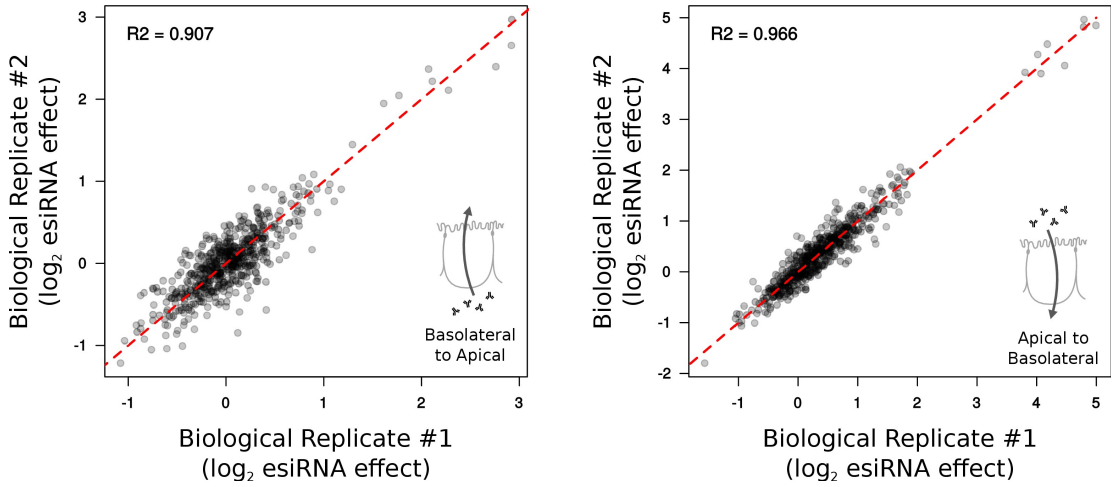


Figure 2.9: Reproducibility of biological replicates during full screen.

This screen was effectively a double screen, with independent data for each of two directions of transport. When we compared Z-scores between the two directions, we discovered that a simple linear fit could entirely explain any difference in Z-score between directions (**Fig 2.10**). This can be explained because we are essentially measuring steady state flux of FcRn across the cell, and at steady state FcRn flux in both directions must be equal. Unfortunately, this means that having data for two directions of transport does not provide any additional biological information, and is useful only as replicates of each other.

We thus pooled together the Z-scores in both directions, resulting in a single pooled Z-score for each esiRNA.

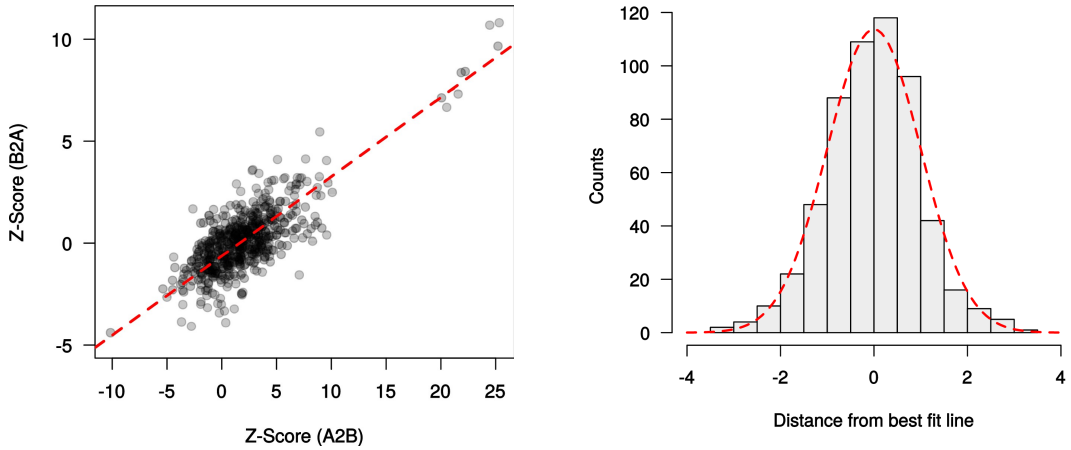


Figure 2.10: Comparison of apical to basolateral (A2B) and basolateral to apical (B2A) transport. Left, Z-scores for A2B and B2A transport for all esiRNAs; dotted red line, the best fit line from a total least squares regression. Right, histogram of the distance from the best fit line. If both directions of transport are fully explained by the best fit line, this histogram is expected to fit a standard normal distribution (mean = 0, sd = 1). The standard normal distribution, plotted as a dotted red line, fits the observed data.

An ordered plot of final Z-scores is shown in **Figure 2.11**. Supporting the validity of our results, 7 out of 9 genes encoding subunits of the exocyst complex resulted in a lower Z-score than all 112 mock transfection controls. The top 15 hits for decreased transport (**Table 2.1**) include both genes with an established role in epithelial membrane trafficking (e.g. Exocyst, *PARD6B*) and genes that have not been linked to epithelial trafficking before (e.g. *ARL14*, *LEPROT*, *C1H6orf211*).

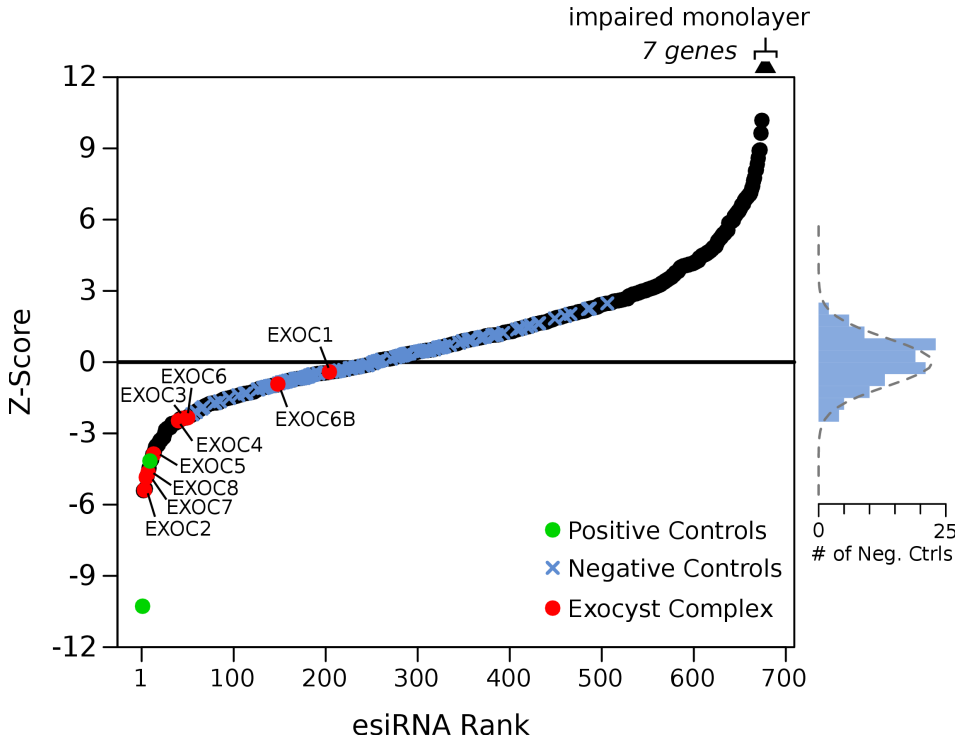


Figure 2.11: Ordered plot of Z-scores for the full screen, with positive controls highlighted in green, negative controls in blue, the exocyst complex in red, and all other esiRNAs in black. A histogram of negative control values is shown on the right, with the standard normal distribution shown as a dotted gray line for comparison.

Table 2.1: Top hits for decreased transport

No.	Gene	Z-Score	Confirmed in 2nd Screen
1*	ACTR3B	-5.41	
2*	EXOC2	-5.38	Y
3*	ARL14	-5.33	Y
4	EXOC7	-4.85	
5*	PARD6B	-4.78	Y
6	EXOC8	-4.64	
7*	VPS13C	-4.47	Y
8*	CCZ1	-4.15	
9	GNAS	-4.08	
10*	LEPROT	-3.89	Y
11	EXOC5	-3.87	
12	STAM	-3.75	
13	DYNC1H1	-3.58	Y
14	DYNC1LI1	-3.52	
15*	ARMT	-3.51	
16	STX19	-3.48	
17	LRRK2	-3.37	
18	RAB10	-3.30	
19	VPS26A	-3.29	Y
20	SNX27	-3.22	
21	TRIP11	-3.19	
22	TMED2	-3.16	
23	ROCK1	-3.01	

*indicates that a gene was selected for small-scale validation

2.4 Confirmation Screen and Small-Scale Validation Assays

To assess the reproducibility of our screen between days, we re-screened nearly 1/3 of the library (164 genes). This confirmation screen included all genes with a Z-score less than -3 or greater than 5, and a selection of genes that were near these Z-score cutoffs. Unfortunately, we observed fairly low reproducibility between screens, with only ~30% of hits displaying a phenotype on the second screen (e.g. only 7 of 23 decreased hits had a Z-score < -2 on the second screen). Thus, although we find very high reproducibility within a day, there is strong variability in esiRNA phenotypes between days.

Due to the variability between the full screen and confirmation screen, we decided to follow up on 8 out of 15 of our top hits in small scale assays where we could more carefully control several parameters (**Table 1**, genes with an asterix). For this follow up, we applied a double transfection protocol that produced stronger and more consistent gene knock down, assessed gene knock down by qRT-PCR, and monitored electrical resistance across the cell layer as a measure of overall cell health. In addition, we repeated all assays on multiple independent days to assess day-to-day variability, and synthesized second esiRNAs that target an separate region of each gene. All esiRNAs led to a gene knock down between 70% - 90% (with the exception of the second esiRNA against *ARL14*) and did not produce a change in electrical permeability (all conditions produced a normal electrical resistance of 120-160 Ωcm^2).

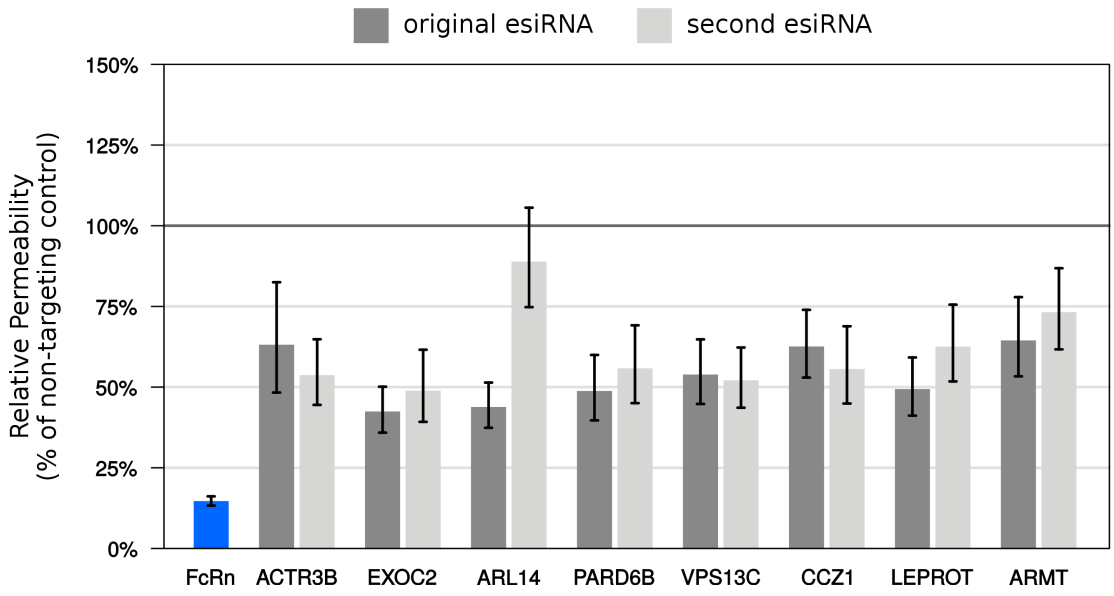


Figure 2.12: Small-scale validation of screen results. Apical to basolateral transport assays for 8 selected hits using two independent esiRNAs. Error bars represent 95% confidence intervals. All experiments reproduced on 2-4 independent days.

In the small scale transport assays, 8 out of 8 original esiRNAs led to a reduction in hGLuc-FcLS permeability that was reproducible between days (**Fig 2.12**), confirming the phenotype from our original screen. In addition, the phenotype was validated for 7 out of 8 genes using second esiRNAs, suggesting that these phenotypes are unlikely to be caused by off-target effects. The second esiRNA against *ARL14* produced a weaker knock down than the original esiRNA, and so the lack of phenotype may be due to insufficient gene knock down. Of the 7 validated hits, only *EXOC2* and *PARD6B* have been previously linked to endosome trafficking in polarized epithelial cells[44,51,52]. Thus, this screen identified several new genes that are necessary for receptor-mediated transcytosis across polarized epithelial cells in culture.

II

Computational Tools to Predict Cell Type-Specific Gene Expression

Unbiased annotation of the regulation, expression and function of mammalian genes requires systematic sampling of the distinct mammalian cell types.

The FANTOM5 Consortium, 2014

3

Estimating Cell Type-Specific Gene Expression through Computational Deconvolution

This chapter describes the difficulties associated with isolating specific cell types for expression analysis, and the potential of computational methods to overcome these difficulties by predicting cell type-specific gene expression without the need for cell isolation. This background information is relevant to the computational method, *CellMapper*, described in Chapter 4 and the resulting applications described in Chapter 5.

3.0 Introduction

The identification of cell-type specific gene expression is key to understanding cellular function and differentiation, and how these processes are disrupted during disease pathogenesis. However, there are steep technical challenges to obtaining pure cell populations for expression profiling[53]. Despite calls for the “systematic sampling” of human cell types[54], current expression resources have predominantly sampled cell types that can

be readily isolated or grown in culture, and expression data is lacking for many important human cell types. ***Computational deconvolution***[55] is a class of methods that can extract cell type-specific information from expression data on heterogeneous cell mixtures. As these methods avoid the expense and technical difficulties of physical purification, they offer the potential to fill important gaps in the expression data. In this chapter, I first provide some background on experimental techniques and resources for cell isolation, and then explain the concept behind computational deconvolution and review existing methods. In Chapter 4, I introduce a new algorithm, ***CellMapper***, that is more robust and sensitive than other approaches, and demonstrate that this algorithm provides accurate predictions for cell types that could not be analyzed previously. Finally, in Chapter 5 I describe applications of CellMapper to (i) prioritize human disease candidate genes and (ii) investigate the expression of a poorly understood cell type called ***enteric glia***.

3.1 Experimental approaches

To measure gene expression in any new cell type, the major technical challenge is often isolating the cell type from its surrounding tissue[53]. Crude cell isolation can sometimes be accomplished by simple mechanical and biochemical protocols involving cell fractionation and basic chemical or enzymatic treatments (e.g. white blood cell fractionation from whole blood, separation of epithelial cell sheets from the intestine). However, most cell types require more advanced methods to purify cells based on specific properties, such as the expression of a marker gene. It can takes years to develop and validate any new protocol for cell isolation, and established methods often require the use of transgenic

animals[56–58] and therefore cannot be applied to humans.

Most experimental isolation methods – including ***fluorescence-activated cell sorting*** (FACS), ***immunopanning***, and ***tandem ribosome affinity purification*** (TRAP) – separate cells based on the expression of a cell-specific marker gene[53]. In preparation for these methods, cells that express the desired marker gene must first be labeled. One approach for cell labeling is to stain a sample with antibodies that recognize a cell-specific epitope; this method requires an antibody to be available for a cell-specific marker gene, and is difficult to apply to intracellular markers. In another approach, a transgenic animal is created in which a reporter gene (e.g. GFP) is expressed using a cell-specific promoter; this method is more effective for intracellular markers but requires the availability of transgenic animals.

When cell-specific marker genes are not available, ***laser capture microdissection*** (LCM) or ***single cell RNA-sequencing*** (single-cell RNA-Seq) can be used. In LCM, a defined region of a biological sample is selected under microscopic observation, and this region is cut out with a laser and separated from the surrounded tissue. In practice, many cell types cannot be identified by morphology alone, and LCM often still requires a cell-specific reporter gene to distinguish the cell population of interest[59]. Single-cell RNA-seq can decompose a tissue into cell populations without requiring any knowledge of the underlying cell types[60], by sequencing RNA from many isolated single cells in parallel and clustering the resulting expression profiles. However, single-cell RNA-seq data is only available for a few organs[60–62]; furthermore, the read densities currently employed by single-cell RNA-seq studies make it difficult to sensitively identify cell type-enriched genes, especially for rare cell types and for genes with relatively low expression levels.

3.2 Computational approaches

3.2.1 Gene-driven differential expression analysis

As described above, the standard approach to identify cell type-enriched genes is to isolate the cell type of interest and search for genes strongly expressed in that cell type relative to a subset of others. However, there is an alternative to this ***sample-driven*** strategy for differential expression analysis that bypasses the need to isolate cell types altogether[63–65]: ***gene-driven*** analysis. Rather than comparing expression between samples, gene-driven analysis searches for genes that share a similar expression profile to an established set of cell type-specific markers, referred to here as ***query genes*** (**Fig 3.1**). This approach can identify cell type-enriched genes using heterogeneous samples such as whole tissue, because the relative proportion of cell types varies from sample to sample. As a result, gene-driven analysis offers the potential to rapidly leverage existing expression data to predict cell type-enriched genes – even if a cell type has not been isolated for expression analysis before.

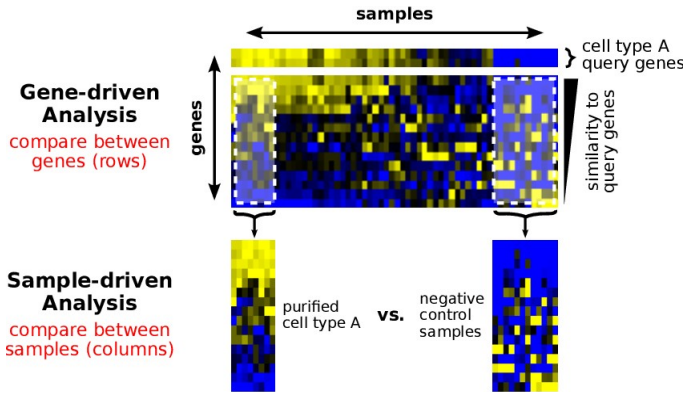


Figure 3.1: Hypothetical example illustrating the difference between sample-driven and gene-driven strategies to identify cell type-specific genes. Sample-driven analysis compares samples (columns) to find genes that are differentially expressed between a purified cell type and negative control samples. Gene-driven analysis compares genes (rows) to find those with a similar expression profile to a set of cell type-specific query genes. Gene-driven analysis can utilize both purified and mixed cell type samples.

Despite its promise, the gene-driven strategy has so far been constrained by limitations of existing methods for analysis. The first application of gene-driven analysis predicted smooth muscle genes using Pearson's correlation[63]. However, correlation was later shown to lack the sensitivity required to identify genes expressed in many other tissues[64] and cell types[65]. To overcome this limitation, more sophisticated algorithms were developed that use machine learning[64–66]. The machine-learning algorithms are effective where correlation fails, but the sensitivity gained comes at a high cost: all require very large training sets of both positive and negative control genes (>20 of each) to define a cell type. This requirement poses a severe limitation for most biological applications, as it is difficult to curate such a large list of established marker genes for even a well-studied cell type, and impossible for many others.

3.2.2 Computational deconvolution and related methods

Gene-driven analysis can be grouped into a larger class of methods called ***computational deconvolution***. Computational deconvolution comprises any method to infer information about individual cell types using expression data from heterogeneous cell mixtures[55]. These methods vary in the resolution of cell-specific information that is extracted from a dataset. At the lower end of resolution are algorithms that estimate only the relative frequencies of each cell type in a sample[55,67], and do not attempt to determine the expression level of any gene. High resolution algorithms attempt ***complete deconvolution***, estimating the absolute expression level of all genes in every (or most) cell types of a sample[68–70]. Gene-driven algorithms provide an intermediate level of

resolution, as they identify genes that are differentially expressed in one cell type relative to others, but do not reveal any information about absolute expression levels. While complete deconvolution provides the most information about a sample, it often requires stronger assumptions and necessitates additional user-provided input. In Appendix B, we show that our gene-driven algorithm, *CellMapper*, strongly outperforms available algorithms for complete deconvolution when absolute expression levels are not needed.

4

Design and Validation of the *CellMapper* Algorithm

This chapter describes the development of *CellMapper*, an algorithm to predict genes selectively expressed in specific cell types. I designed and conducted all experiments discussed in this chapter, with general strategic guidance from my advisor, Wayne Lencer, and support on the computational biology and statistics from our collaborators Curtis Huttenhower and Levi Waldron.

4.0 Introduction

Cell type-specific gene expression plays a defining role in cellular function and differentiation. However, efforts to compare gene expression across cell types have been confounded by the challenge of isolating pure cell populations and the cellular heterogeneity of mammalian tissues. The human brain provides a clear example: many brain cell types do not maintain their differentiated state when grown in culture, and can only be isolated acutely from intact brain tissue. Validated cell isolation protocols in mice

often require the use of transgenic animals to label specific cell types[56–58,71,72], and are therefore not applicable to human. As a result, expression data are only available for a small fraction of the ~150 estimated cell types[73] of the human central nervous system, and this problem is similar for many other tissues.

Gene-driven analysis (reviewed in Chapter 3) introduces an appealing alternative to cell isolation: expression profiles of individual cell types can be computationally inferred from heterogeneous cell mixtures, avoiding the expense and technical difficulties of physical purification. Rather than comparing expression between samples, gene-driven analysis searches for genes with a similar expression profile to known cell type-specific markers, called “query genes”. The gene-driven strategy has been used to identify genes expressed selectively in smooth muscle cells[63] and kidney podocytes[65], demonstrating that it can match or exceed the accuracy of targeted expression profiling studies, and is effective even for cell types that are difficult to isolate. However, this approach has so far been constrained by limitations of existing methods for analysis; in particular, the most current and sensitive algorithms require very large training sets of positive and negative control genes to define each cell type[65], limiting their application to cell types where many marker genes are already available.

Here, we describe an approach to substantially increase the sensitivity of gene-driven analysis, making it possible to rapidly and accurately predict cell type-specific genes for almost any cell type – even when only a single query gene is available. The main innovation is a singular value decomposition (SVD) filter that serves to highlight subtle, but biologically relevant signals in the data. We then apply our algorithm to a large compendium of 19,801 microarrays and identify genes specifically expressed in 30

diverse cell types of widespread importance in human biology. Our approach can be applied to any transcriptionally defined cell population.

4.1 Development of CellMapper

4.1.1 Comparing prospective algorithms

To establish a gene-driven algorithm that is accurate with only a small number of known marker genes (1-2 genes), we started by comparing several prospective algorithms against a benchmark of genes enriched in each of 30 tissues[74]. We initially focused on tissue-specific gene expression (e.g. liver, intestine, heart), rather than cell type-specific expression, because there exist large catalogs of tissue-specific genes to serve as a “gold standard” for performance evaluation. This strategy also allowed us to perform algorithm development and optimization using an independent test case (tissue-specific expression), and fix all algorithm parameters before moving on to our primary interest (cell type-specific expression).

Pearson's correlation has been applied to identify cell type-enriched genes using a single query gene in a few cases[63,75], but was later shown to lack the sensitivity required for many tissues[64] and cell types[65]. To overcome this limitation, we first tested several algorithms that were originally developed to find genes in co-regulated biological pathways (e.g. genes associated with the same GO terms) – GeneRecommender[76], MEM[77], and SPELL[78] – as well as mutual information. Each of these algorithms are compatible with small training sets (1-2 query genes) and have greater sensitivity than Pearson's correlation when applied to biological pathways; we hypothesized that one of these

alternative algorithms might also be effective when applied to cell types.

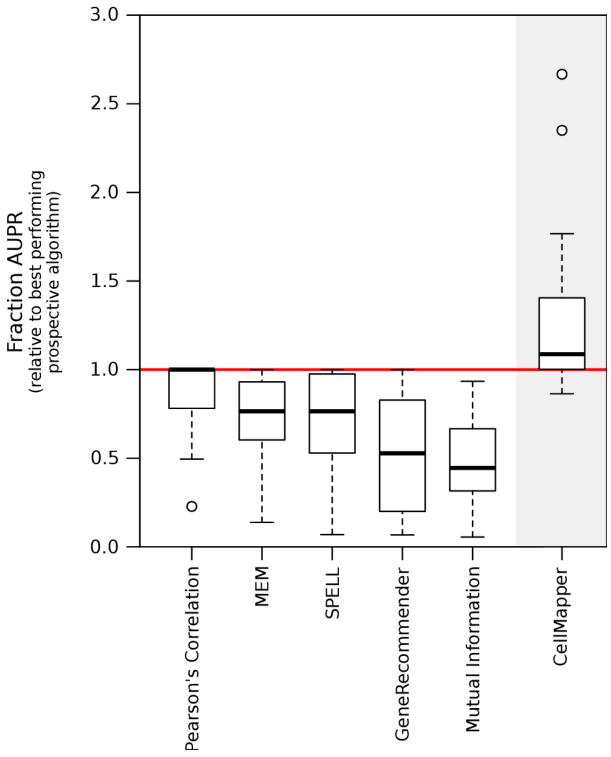


Figure 4.1: Performance evaluation of five prospective gene-driven search algorithms using TiGER tissue genes as a gold standard[74], compared to the final algorithm CellMapper. Tukey boxplots show the change in area under the precision recall curve (AUPR) for each tissue, relative to the AUPR achieved by the best-performing prospective algorithm for that tissue. While all five prospective algorithms performed poorly relative to the other algorithms in several tissues, CellMapper achieved the highest AUPR in 25 out of 30 tissues and was always within 20% AUPR of the best method. MEM, Multi Experiment Matrix[77]; SPELL, Serial Patterns of Expression Levels Locator[78]; GR, Gene Recommender[76]; MI, Mutual Information.

To compare algorithms, we applied each to search a large compendium of human microarrays[79] using many independent combinations of 2 query genes for each tissue. After each search, we then quantified the accuracy with which the remaining TiGER genes from the same tissue were identified (leave-2-in cross validation) as assessed by the area under the precision-recall curve (AUPR). Unfortunately, none of the newer algorithms

provided a consistent performance increase relative to Pearson's correlation. While each strongly outperformed correlation in some tissues, they all performed very poorly in many others (**Fig 4.1**). Overall, the relative performance of the five algorithms was highly variable between tissues, with no single algorithm performing well across the board. This lead us to test alternative strategies to increase the sensitivity of Pearson's correlation, and we found success when filtering the data based on singular value decomposition (SVD), as described below.

4.1.2 SVD filter design and optimization

In our comparison across algorithms, we found that Pearson's correlation performed best on average, but very poorly for some tissues (**Fig 4.1**). To address this limitation, we sought a strategy to harvest more information from the data without requiring an increase in the number of query genes. One approach to highlight subtle, but informative signals in microarray data is to filter the data based on singular value decomposition[80] (SVD). Singular value decomposition (SVD; also related to principal component analysis) of an expression matrix is the linear transformation of the original m genes by n arrays into an uncorrelated set of “eigengenes” and “eigenarrays” [80] given by:

$$X_{m\times n} = U_{m\times n}\Sigma_{n\times n}V^{T}_{n\times n}$$

where X is the expression matrix; U and V contain the eigenarrays (right-singular vectors) and corresponding eigengenes (left-singular vectors) of X, respectively; and Σ contains the singular values of X, or the relative importance (variance explained) of each eigenvector in the original expression matrix. SVD is widely used in genomics data

analysis because the eigengenes and eigenarrays often have a biological interpretation. For instance, in the HG-U133A dataset used in this study, the first eigengene distinguishes hematopoietic from solid tissue samples[79], and the first eigenarray explains the corresponding genomic expression changes that accompany hematopoiesis.

While the top eigenvectors represent the strongest signals from the original expression matrix, they are not the most informative for every biological question. For instance, in an SVD analysis of yeast cell cycle microarrays, the first eigenvector explained over 90% of the gene expression data, yet the second and third eigenvectors contained most of the oscillating cell cycle gene expression signal[80]. The first eigenvectors can also relate to systematic technical noise such as lab effects[80,81]. Finally, the strongest signals in a large meta-analysis of diverse samples will be dominated by the types of experiments performed most often in the literature; almost a third of the HG-U133A dataset contains microarrays from breast or breast cancer[79]. This sampling bias will disproportionately impact the first eigenvectors, while later eigenvectors may contain relevant information from biological conditions sampled less frequently. To increase the influence of potentially informative signatures from the later eigenvectors, we filtered the data by adjusting the relative weight of each eigenvalue.

One possibility would be to posit that each eigenvector has an equal chance of being informative, and weight all eigenvectors equally. The gene co-expression algorithm SPELL effectively takes this approach[78], by examining correlations between genes in eigenarray space. However, earlier eigenvectors contain a greater signal to noise ratio, and so weighting them equally with the lower (and noisier) eigenvectors may result in overemphasis of noise in the later eigenvectors. Therefore, we examined filters of the

form:

$$\sigma_k' = \sigma_k^{\alpha}$$

which varies smoothly between no filter (α = 1) to completely equalized eigenvalues (α = 0; comparable to SPELL). **Figure 4.2A** shows how AUPR varied as a function of α for each of the TiGER tissues. The vast majority of tissues showed an increase in AUPR for most values of α, and many demonstrated an increase in AUPR even as α approached 0. We selected α to be 0.5 because this resulted in an improved AUPR for 25 out of 30 tissues ($p = 3.5 \times 10^{-7}$, Wilcoxon signed rank test), and never led to a substantial decrease.

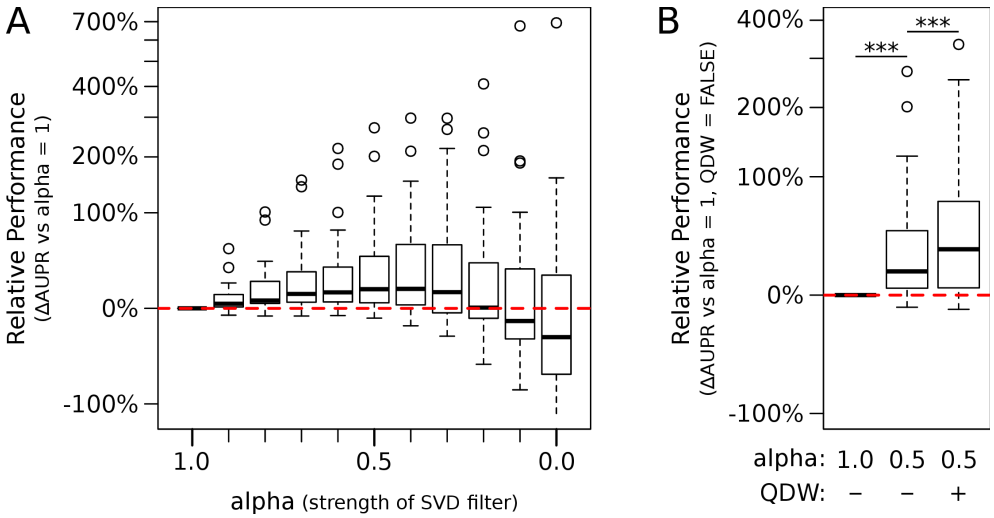


Figure 4.2: Parameter optimization for the CellMapper singular value decomposition (SVD) filter, using test searches to find tissue-enriched genes as defined in the TiGER database[74]. (A) Evaluation of the free parameter, alpha. The SVD filter incorporates a free parameter, alpha, which allows the strength of the filter to be tuned, ranging in value from 1 (weak filter) to 0 (strong filter). Alpha values between 1 and 0.3 led to an increase in AUPR for 25 out of 30 tissues. An intermediate value of 0.5 was chosen for the final algorithm, and this parameter was fixed for all analyses on cell types. (B) Evaluation of the query-driven weight term (QDW). The SVD filter also includes a term, abbreviated QDW, that decreases the weight of components in which the query genes are not well separated from the rest of the genome. The QDW term leads to an increase in performance beyond what is seen using the alpha scaling factor alone. ***, $p < 10\text{-}4$; Wilcoxon singed rank test. In both subfigures, AUPR was plotted relative to alpha = 1 and no query-driven weight term, which is approximately equivalent to Pearson's correlation (it is equal to Pearson's correlation with the low variance principle components filtered, see Methods).

The above filter assumes that there is no way to identify which eigenvectors will best distinguish genes expressed in a given cell type. However, as we are defining cell type genes based on their similarity to a set of query genes, we can expect that the most informative eigenvectors will be those where the query genes are well separated from the rest of the genome. Therefore, we also apply a soft filter to the eigenvectors, multiplying each eigenvalue by a weight that increases as the query genes stand out:

$$w_k = \sum_{g \in (query\ genes)} \tanh\left(u_k^g\right)$$

where u_k^g is the loading of gene g in singular vector k, normalized so that u_k has a mean of 0 across all genes with a standard deviation of 1. This weight plateaus when the query genes are at least a standard deviation away from the mean value for an eigenvector, but approaches 0 as the query genes tend towards the mean. **Figure 4.2B** shows that this query-driven weighting (QDW) produced an increase in AUPR beyond what was obtained using the α filter alone ($p = 9.3 \times 10^{-4}$ relative to the filter with α = 0.5 and no QDW; Wilcoxon signed rank test).

An important benefit of this filter was to make the final algorithm, *CellMapper*, consistent across tissues. While the other algorithms performed inconsistently, CellMapper was always among the best-performing methods, and outperformed every other algorithm in 25 out of 30 tissues (**Fig 4.1**, right). In addition, the SVD filter of CellMapper is robust to added Gaussian noise (**Fig 4.3**) and makes the algorithm less sensitive to bias in sample composition (**Fig 4.4**). After establishing these two suitable filters for ranking of tissue-specific genes, the same filters were applied to the identification of cell-type specific genes.

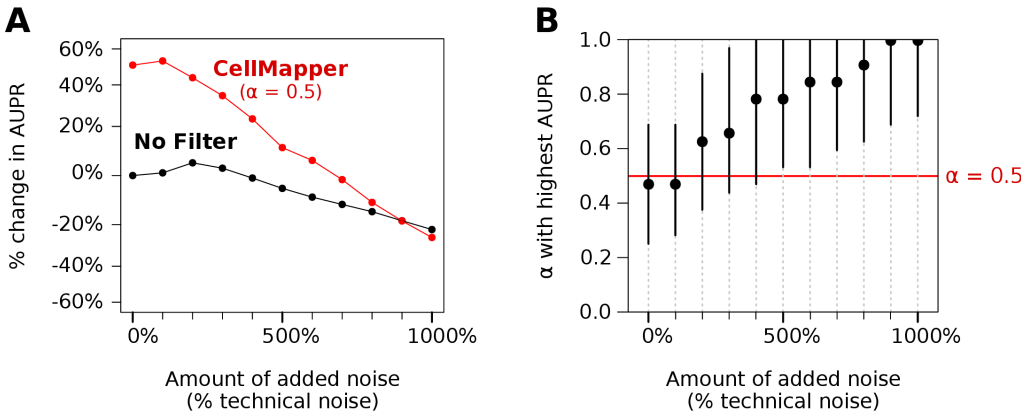


Figure 4.3: Sensitivity of the CellMapper SVD filter to added Gaussian noise. Gaussian noise was added to the Lukk, et al. (2010) dataset through the range of 0% to 1000% of the estimated technical noise, and then CellMapper was applied to search the noisy dataset to find TiGER tissue genes. The amount of technical noise was estimated from the median standard deviation of the 500 least expressed genes. (A) Performance of CellMapper in the face of added technical noise, with and without the SVD filter applied ("No Filter" is equivalent to Pearson's correlation with the low variance principal components removed). AUPR was plotted relative to "No Filter" with 0% added technical noise. At the point where performance with and without the SVD filter is equal (900% added technical noise), the added Gaussian noise has a greater variance than 99% of genes in the dataset. (B) Sensitivity of the choice of free parameter, α, to added noise. The choice of α that results in the maximum AUPR is shown as a black dot, and all choices that result in an AUPR within 5% of this maximum value are covered by the black line. At the point where α = 0.5 (RED line) is no longer within 5% AUPR of the best performing value (500% added technical noise), the added Gaussian noise has a greater variance than 90% of genes in the dataset. These values were calculated with CellMapper's query-driven weight term included, and so α = 1 is not the same as "No Filter". When the query driven weight term was not included, α was less sensitive to noise.

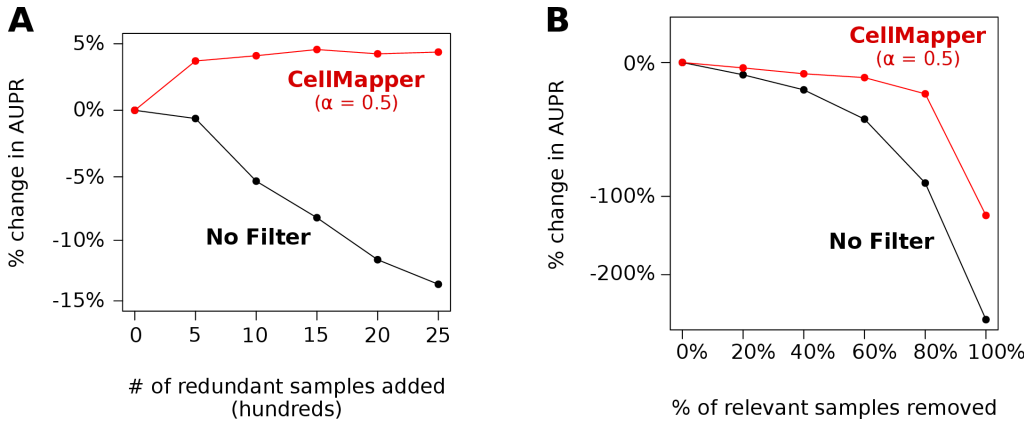


Figure 4.4: Sensitivity of the CellMapper SVD filter to sampling bias. Samples were drawn from the Lukk, et al. (2010) dataset in order to intentionally increase or decrease bias in sample composition, and the affect on algorithm performance was quantified. (A) Sensitivity to adding redundant samples. CellMapper was applied, with and without the SVD filter, to search for tissue-specific genes using 500 randomly selected samples from the total microarray dataset, plus varying numbers of added “redundant samples”. For this analysis, “redundant samples” were selected from a subset of the data annotated as “blood”, “bone marrow”, and “mammary gland” because these 3 sample annotations are the most over-represented in the Lukk dataset, accounting for over half of all samples. While performance degraded when redundant samples were added without the SVD filter, CellMapper actually performed better and was able to benefit from the increase in sample size. (B) Sensitivity to removing relevant samples. Samples annotated as belonging to a specific tissue were removed from the Lukk dataset, and CellMapper was applied to search this truncated dataset for genes expressed in the tissue with samples removed. This analysis was run separately for each of 7 tissues (“bone”, “colon”, “kidney”, “liver”, “ovary”, “prostate”, and “skin”), and the mean change in AUPR across all tissues is reported. These tissues were analyzed because they represent an intermediate number of samples in the Lukk dataset (50-150 sample for each tissue, or 1-3% of the total).

4.2 CellMapper can distinguish cell type-specific expression signatures from whole tissue microarray data

4.2.1 Application to major brain cell classes

As a first application of CellMapper, we applied it to identify genes expressed in four major brain cell types – neurons, astrocytes, oligodendrocytes, and microglia – using microarray data from the Allen Brain Atlas[82]. The Brain Atlas data provides an excellent case study for gene-driven analysis because it is composed exclusively of samples from heterogeneous brain tissue, drawn from many regions with varying cellular composition. This dataset has also been shown to contain sufficient signal to differentiate cell type-specific expression; when it was analyzed using an exploratory clustering method that groups genes into co-expressed modules, several of the resulting modules corresponded to genes expressed in specific cell types[82]. In addition, there are separate experimental datasets from purified samples of all four brain cell types[56–58,83], providing a means for comparison and validation.

We modeled our analysis after a recent RNA-Seq study of gene expression in cell types from the mouse cerebral cortex[58]. In this study, neurons, astrocytes, oligodendrocytes, and microglia were identified by their expression of *L1CAM*, *ALDH1L1*, *MOG*, and *PTPRC* (respectively) and isolated using fluorescence activated cell sorting (FACS) and immunopanning. We applied CellMapper to replicate a similar experiment *in silico*: cell type expression profiles were defined based on the query genes *L1CAM*, *ALDH1L1*, *MOG*, or *PTPRC*, and then other genes with a correlated expression profile were identified. This analysis returned 213 genes for neurons, 474 for astrocytes, 1027 for oligodendrocytes, and 216 for microglia at a false discovery rate (FDR) of 0.01.

To evaluate the accuracy of our results, we took two complementary approaches. In the first, we examined CellMapper predictions for literature-defined markers (positive controls) of each cell type. As positive controls, we selected the cell-specific markers used for validation in three previous studies[57,58,84]. CellMapper correctly associated 20 out of 21 positive control genes with the expected cell type (**Fig 4.5**). The only exception, *GFAP*, is expressed at variable levels within astrocytes[85], and so it is possible this gene was missed for this reason. In the second approach, we asked whether CellMapper predictions for each cell type were enriched for genes associated with these cell types as measured by expression profiling[56–58]. We found that our predictions for neurons, astrocytes, oligodendrocytes, and microglia were significantly enriched for genes expressed in those cell types as measured by the previous RNA-Seq study[58] ($p = 6.3 \times 10^{-69}$, $p = 4.5 \times 10^{-76}$, $p = 2.5 \times 10^{-131}$, and $p = 1.1 \times 10^{-84}$, respectively; Fisher's exact test), and by three other studies[56,57,83] (**Table 4.1**). In contrast, we never observed significant overlap between CellMapper predictions for one cell type and experimentally measured genes from a different cell type (enrichment p-value > 0.05), confirming the specificity of our results.

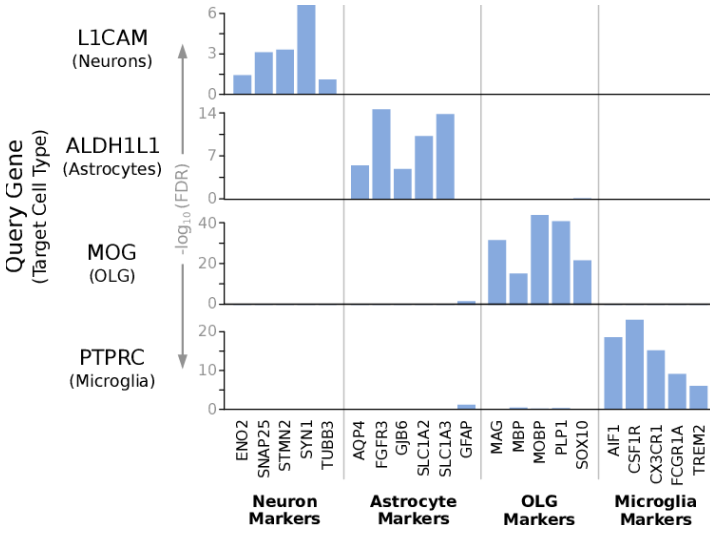


Figure 4.5: CellMapper was applied using query genes for four major brain cell types. The correct identification of classic cell-specific markers confirms the accuracy of CellMapper's predictions. OLG, myelinating oligodendrocytes; FDR, false discovery rate.

Table 4.1. CellMapper Predictions are Enriched for Experimentally Defined Cell Type Genes

Target Cell Type	Query Gene	Neurons[58]	Neurons[57]	Neurons[56]	Astrocytes[58]	Astrocytes[57]	Astrocytes[56]	Oligodendrocytes[58]	Oligodendrocytes[57]	Oligodendrocytes[56]	Microglia[58]	Microglia[83]
Neurons	*L1CAM*	108/198 $p = 6.3 \times 10^{-69}$	99/140 $p = 3.6 \times 10^{-49}$	50/125 $p = 2.5 \times 10^{-23}$	NS	NS	NS	NS	NS	NS	NS	NS
	SNAP25	120/237 $p = 7.5 \times 10^{-72}$	121/160 $p = 3.0 \times 10^{-66}$	65/165 $p = 3.5 \times 10^{-30}$	NS	NS	NS	NS	NS	NS	NS	NS
	STMN2	160/417 $p = 2.5 \times 10^{-74}$	152/255 $p = 2.3 \times 10^{-61}$	86/277 $p = 2.9 \times 10^{-31}$	NS	NS	NS	NS	NS	NS	NS	NS
Astrocytes	*ALDH1L1*	NS	NS	NS	141/437 $p = 4.5 \times 10^{-76}$	196/292 $p = 9.2 \times 10^{-117}$	155/320 $p = 2.3 \times 10^{-99}$	NS	NS	NS	NS	NS
	FGFR3	NS	NS	NS	216/681 $p = 1.0 \times 10^{-118}$	288/441 $p = 3.6 \times 10^{-174}$	203/480 $p = 6.6 \times 10^{-119}$	NS	NS	NS	NS	NS
	SLC1A3	NS	NS	NS	169/494 $p = 5.4 \times 10^{-97}$	226/323 $p = 7.4 \times 10^{-143}$	157/351 $p = 3.5 \times 10^{-94}$	NS	NS	NS	NS	NS
Oligodendrocytes	*MOG*	NS	NS	NS	NS	NS	NS	186/929 $p = 2.5 \times 10^{-131}$	209/492 $p = 1.2 \times 10^{-109}$	88/596 $p = 1.8 \times 10^{-57}$	NS	NS
	MOBP	NS	NS	NS	NS	NS	NS	181/825 $p = 9.3 \times 10^{-135}$	195/436 $p = 2.8 \times 10^{-106}$	84/520 $p = 1.1 \times 10^{-57}$	NS	NS
	PLP1	NS	NS	NS	NS	NS	NS	179/847 $p = 6.6 \times 10^{-130}$	204/445 $p = 1.2 \times 10^{-114}$	88/546 $p = 1.3 \times 10^{-60}$	NS	NS
Microglia	*PTPRC*	NS	NS	NS	NS	NS	NS	NS	NS	NS	123/192 $p = 1.1 \times 10^{-84}$	21/216 $p = 5.1 \times 10^{-16}$
	AIF1	NS	NS	NS	NS	NS	NS	NS	NS	NS	238/492 $p = 3.1 \times 10^{-129}$	35/564 $p = 4.4 \times 10^{-21}$
	CX3CR1	NS	NS	NS	NS	NS	NS	NS	NS	NS	184/316 $p = 1.4 \times 10^{-117}$	32/355 $p = 3.4 \times 10^{-24}$

Enrichment of experimentally defined cell type genes within CellMapper predictions. Superscript numbers indicate the reference from which data were taken. Fractions are the number of overlapping genes divided by the maximum possible number of overlapping genes. NS, not significant ($p > 0.05$).

These findings were robust to the choice of query gene, as both the literature-curated markers and experimentally-defined cell type genes were also correctly identified when CellMapper was run using different query genes (**Table 4.1**). Thus, CellMapper accurately identified genes expressed selectively in these four cell types.

4.2.2 Application to other cell types

We next tested CellMapper on a large panel of additional cell types (**Table 4.2**), this time extending our analysis to include both brain and non-brain cell types, with multiple representatives of all major cell classes (neural, epithelial, connective tissue, muscle, and hematopoietic). We curated one query gene for each cell type (**Table 4.2**); in most cases, these query genes were selected because their promoters are used to drive cell type-specific Cre expression in validated conditional mouse knock out models[86]. Then we applied CellMapper to search our microarray datasets using the query genes, predicting a mean of 371 cell type-enriched genes per cell type (FDR ≤ 0.01). Again, the quality of our results was evaluated using literature-curated positive control genes as well as a set of negative control genes, which include cell-specific markers for non-target cell types and a reference set of housekeeping genes[87]. For every cell type, CellMapper identified over half of the positive control genes within the top 100 predictions (**Fig 4.6**), and excluded almost every negative control gene. In total, 208 out of 241 positive controls were ranked within the top 100 predictions for the correct cell type (86.3%), and all but two were ranked within the top 516 predictions (99.2%; the only exceptions were *GFAP* within astrocytes, discussed above, and *SYP* within enteroendocrine cells). Thus, CellMapper is accurate for

both single- and multi-organ cell types, and for cell types difficult to isolate or culture (e.g. Schwann cells, Paneth cells).

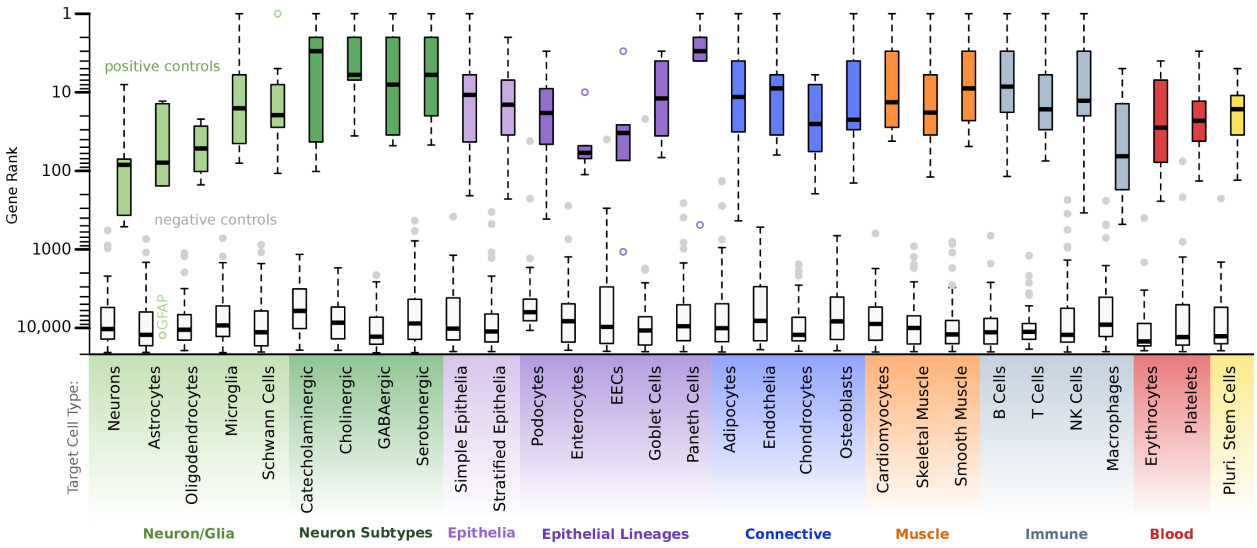


Figure 4.6: CellMapper is accurate across diverse cell types. CellMapper was applied using query genes for 30 cell types (**Table 4.2**); Tukey boxplots display the rank of 5-10 literature curated markers (positive controls) and ≥48 negative control genes for each cell type, demonstrating that CellMapper sensitively identified established cell type-specific markers in every case. Filled gray circles represent negative control genes that fall outside 1.5 times the inter quartile range of the other negative control genes (“outliers”); open circles represent positive control genes that fall outside this range. In only four instances (0.3%) was a negative control gene identified within the top 100 predictions for a cell type. EECs, enteroendocrine cells.

Table 4.2. Cell Types and Query Genes.

Cell Type	Query Gene	Experiment Group	Microarray Dataset(s)
Neurons	*L1CAM*	A	Allen Brain Atlas
Astrocytes	*ALDH1L1*	A	Allen Brain Atlas
Oligodendrocytes (Myelinating)	*MOG*	A	Allen Brain Atlas
Microglia	*PTPRC*	A	Allen Brain Atlas
Catecholaminergic Neurons	*TH*	B	Allen Brain Atlas
Cholinergic Neurons	*CHAT*	B	Allen Brain Atlas
GABAergic Neurons	*GAD1*	B	Allen Brain Atlas
Serotonergic Neurons	*FEV*	B	Allen Brain Atlas
Adipocytes	*FABP4*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
B Cells	*CD19*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Cardiomyocytes	*TNNI3*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Chondrocytes	*ACAN*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Endothelial Cells	*TEK*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Erythrocytes	*EPB42*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Macrophages	*CD163*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
NK Cells	*NCR1*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)

Osteoblasts	*IBSP*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Platelets	*PF4*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Pluripotent Stem Cells	*NANOG*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Schwann Cells (Myelinating)	*MPZ*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Simple Epithelial Cells	*KRT8*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Skeletal Muscle Cells	*TNNT3*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Smooth Muscle Cells	*MYH11*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Stratified Epithelial Cells	*KRT5*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
T Cells	*CD3D*	C	Engreitz, et al. (2010); Lukk, et al. (2010); Zheng-Bradley, et al. (2010)
Enterocytes	*ALPI*	D	Intestine-specific subset of Engreitz, et al. (2010) and Lukk, et al. (2010)
Enteroendocrine Cells	*CHGA*	D	Intestine-specific subset of Engreitz, et al. (2010) and Lukk, et al. (2010)
Goblet Cells	*MUC2*	D	Intestine-specific subset of Engreitz, et al. (2010) and Lukk, et al. (2010)
Paneth Cells	*DEFA5*	D	Intestine-specific subset of Engreitz, et al. (2010) and Lukk, et al. (2010)
Podocytes	*PTPRO*	E	Kidney Datasets from Ju, et al. (2010)

Table 1 (cont.). There were five “Experiment Groups”, each of which used a different set of microarray data and different control genes (control genes are negative control markers supplied to CellMapper, see Methods). As control genes, we included all query genes for a non-target cell type within the same experiment group, with the addition of ALDH1L1, MOG, and PTPRC for Group B and AQP1, ACTA2, and CDH5 for Group E.

4.2.3 Experimental validation of novel predictions

As a final test of the gene-driven approach, we experimentally validated three new predictions by RNA in situ hybridization (ISH). We focused on predictions for smooth muscle and simple epithelial cells because these cell types are in close proximity in many tissues but can be identified without co-staining for cell-specific markers. **Figures 4.7** shows the expression of two poorly studied genes newly predicted for simple epithelia (*TMEM30B* and *C77080*) and one for smooth muscle (*RASL12*) in six different mouse tissues. We found that *TMEM30B* and *C77080* were expressed in the simple epithelial cells of every tissue examined, with no detectable staining in fibroblasts, muscle cells, endothelial cells, or other connective tissue. *RASL12* was expressed strongly in vascular smooth muscle from every tissue and smooth muscle from the epididymus, weakly in the external smooth muscle layers lining the colon, but absent from connective tissue and epithelia. We also observed *C77080* and *RASL12* expression in a subset of cells within autonomic ganglia located near the heart (**Fig 4.7E**), the only example of an additional site of expression. Taken together, all three genes were strongly enriched in the predicted cell type relative to a wide range of others across 6 diverse mouse tissues.

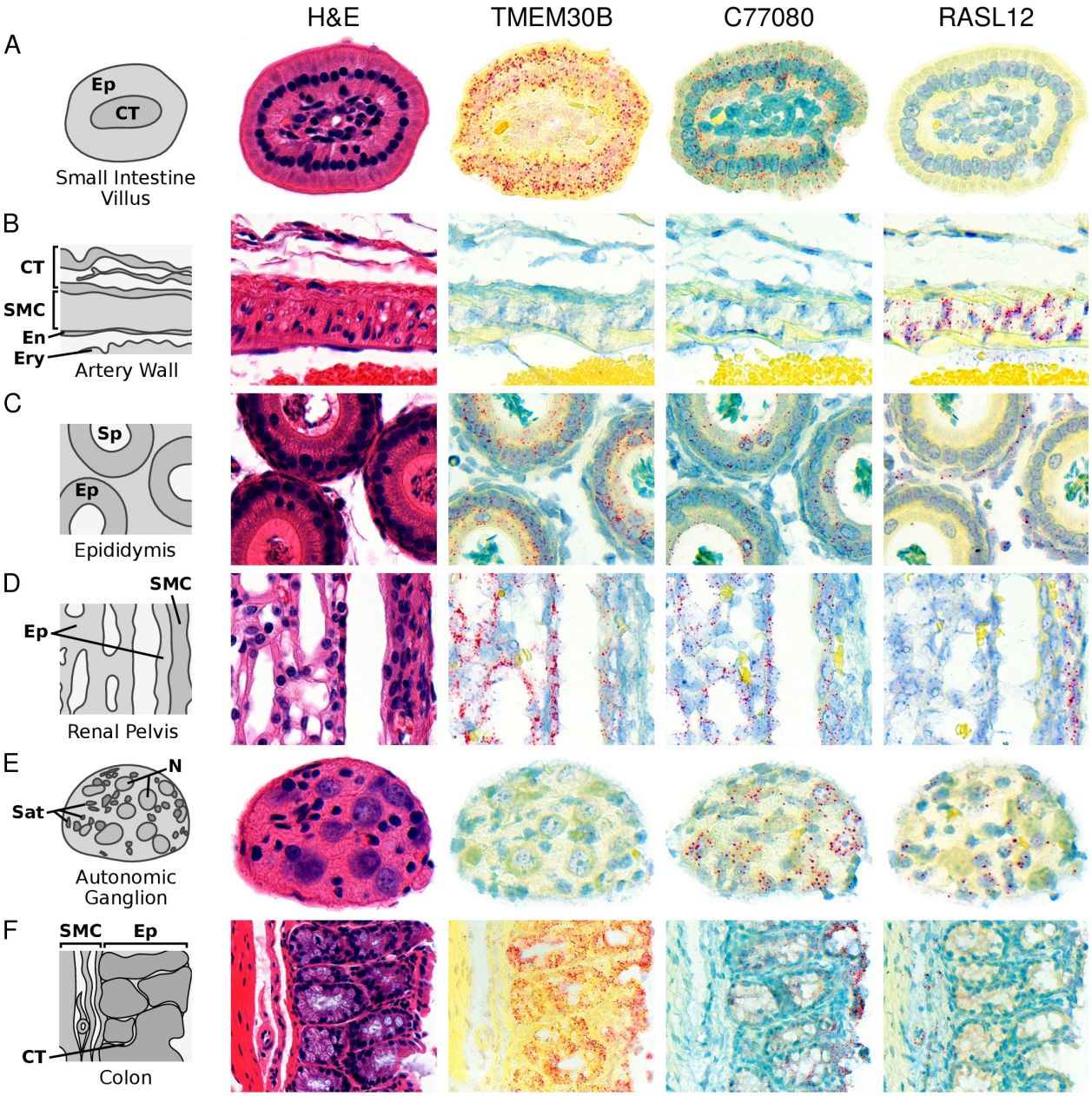


Figure 4.7: Validation of novel predictions. RNA in situ hybridization of serial sections of selected mouse tissues using probes against *TMEM30B*, *C77080*, or *RASL12*. Red dots indicate staining with Affymetrix QuantiGene ViewRNA probes, blue is hematoxylin counterstain. Epi, epithelial cells; SMC, smooth muscle cells; N, neuron cell bodies; Sat, satellite cells. (A) Cross section of villus from small intestine. (B) Wall of medium sized artery sectioned longitudinally, with connective tissue (above) and lumen (below). (C) Epididymus, including epithelium, and underlying connective tissue rich in smooth muscle cells. (D) Renal pelvis, including edge of medulla with collecting ducts (left) and urinary epithelium of calyx with associated smooth muscle (right). (E) Autonomic ganglion containing neurons

Figure 4.7 (cont.): and satellite cells. (F) Colon, including epithelium of mucosal surface (right) and crypts (center), muscularis mucosae under crypt bases, submucosa with small arteriole, and circular smooth muscle layer (left).

4.2.3 Comparison to related approaches

A gene-driven algorithm, called *in silico* nano-dissection, was recently shown to be effective at identifying genes expressed in kidney podocytes[65]. This algorithm is based on machine learning and requires a large training set of positive and negative control genes. To test how the accuracy of CellMapper compared to *in silico* nano-dissection, we applied each method to identify genes expressed in podocytes and quantified the accuracy with which each recovered an experimentally defined ("gold standard") set of podocyte genes[88]. We found that CellMapper identified the

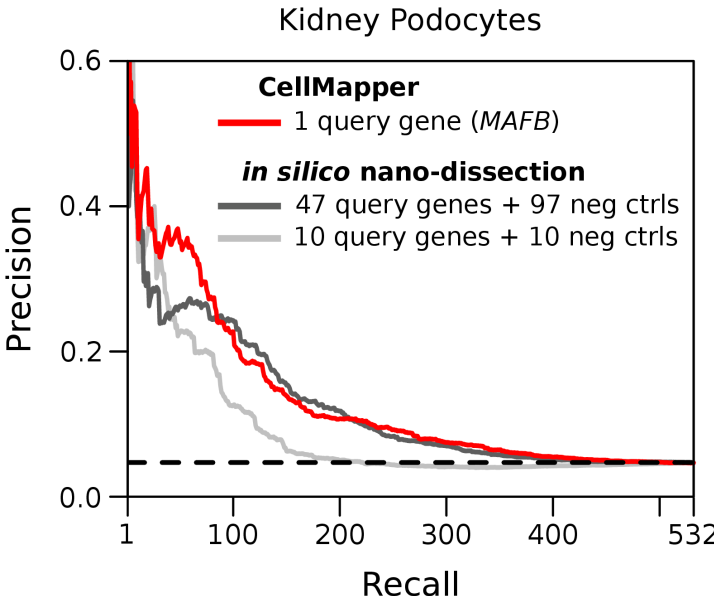


Figure 4.8: CellMapper achieves similar accuracy to the machine learning algorithm, *in silico* nano-dissection[65], while requiring far fewer query genes. Evaluating in silico nano-dissection and CellMapper based on the ability to recover an independent experimentally defined set of podocyte-enriched genes[88]. Dark gray line, *in silico* nano-dissection using the training set from the original study (46 query genes and 97 negative control genes); light gray line, *in silico* nano-dissection using a smaller set of 10 query genes and 10 negative control genes, the smallest training set permitted by the algorithm. CellMapper achieved comparable precision to in silico nano-dissection using a substantially smaller training set (1 query gene versus 46 for in silico nano-dissection).

experimentally defined podocyte genes with similar precision to *in silico* nano-dissection at all levels of recall (**Fig 4.8**), despite using a much smaller training set of positive and negative control genes (1 query gene for CellMapper vs. 46 query genes and 97 negative controls for *in silico* nano-dissection). We then repeated *in silico* nano-dissection using a smaller training set of 10 query genes and 10 negative control genes (the smallest training set permitted by the algorithm), choosing ten established podocyte markers as query genes (*CR1*, *MAFB*, *MME*, *NES*, *NPHS1*, *NPHS2*, *PDPN*, *PODXL*, *TJP1*, and *WT1*) and markers for the other major kidney cell types as negative control genes (negative controls: *CDH5*, *KDR*, and *TEK* for endothelia; *ACTA2*, *CD34*, and *PDGFRB* for mesangial cells; *AQP1*, *SLC12A1*, *SLC12A3*, and *UMOD* for tubule cells). When using this smaller training set, we observed a decrease in performance for *in silico* nano-dissection, such that it performed noticeably worse than CellMapper (**Fig 4.8**, light gray line). Thus, CellMapper achieved similar accuracy to *in silico* nano-dissection while requiring substantially fewer query genes.

Large training sets of positive and negative control genes are not available for many biologically important cell types (e.g. the neuronal and intestinal epithelial subtypes we analyzed), and so the ability to use a single query gene was essential to the success of our analysis. A more detailed comparison of CellMapper to other computational approaches is provided in Appendix B; all other approaches have limitations that prevent their application to many of the 30 cell types we analyzed successfully with CellMapper.

4.3 Discussion

Understanding the unique gene expression profiles of individual cell types is fundamental to continued advances in biology and medicine. CellMapper is one approach to obtain this crucial information. We show the method to be sensitive, robust, and highly capable of addressing both basic and clinical problems in human biology.

Perhaps the most important advance is that CellMapper can be applied to many cell types that would be difficult or impossible to approach with other methods. Earlier gene-driven algorithms performed inconsistently or required very large training sets of positive and negative control genes. CellMapper, in contrast, maintains high sensitivity and specificity when using only a single marker gene, a condition critical for enabling many biological applications. Compared to sample-driven experimental studies, CellMapper can be applied to cell types that are difficult to purify. For example, many of the brain cell types we investigated – such as myelinating oligodendrocytes and several neuron subtypes – have not been isolated for expression analysis from humans before. Protocols to isolate these cell types from mice required the use of transgenic animals and other reagents not available for application to humans[56–58]. However, with CellMapper we were able to predict genes expressed in these cell types using human microarray data from complex brain tissue.

We also emphasize the practical ease with which gene-driven analysis by CellMapper can be applied. While sample driven approaches require a substantial investment in time and resources to purify or enrich each cell type, CellMapper requires only a single marker gene and readily available microarray data. Markers can be used to

delineate not only individual cell lineages (*DEF5A*+ Paneth cells), but also larger classes of cells with similar function (*KRT8*+ simple epithelia), allowing the level of resolution to be tailored to the needs of each specific biological question. This makes it feasible to rapidly and accurately define genes expressed in many cell types in parallel, as we have demonstrated for 30 widely diverse cell types.

A built-in limitation of gene-driven approaches, such as CellMapper, is that they are dependent on the availability of cell-specific marker genes and large, representative expression datasets. Fortunately, marker genes have been established for a wide variety of cell types, and the requirement of a single marker gene is no greater than that needed by experimental techniques such as FACS and immunohistochemistry. The availability of expression data will be most limiting for rare cell types that populate a single tissue, but we show that CellMapper can still separate genes expressed in closely related cell types such as neuron subtypes and intestinal epithelial lineages. Another limitation of our approach is that it only addresses genes covered by micoarrays; certain classes of genes, such as long non-coding RNAs, are not well represented and so CellMapper cannot make predictions for these genes. Future work could adapt the method for RNA-Seq data to allow for more complete coverage of the transcriptome. By enabling gene-driven analysis to a broader range of cell types, CellMapper allows for diverse applications in biology and medicine.

5

Applications of CellMapper

This chapter describes two examples where the cell type-expression predictions of CellMapper can shed insight into specific biological problems. In the first example, we applied CellMapper to prioritize candidate genes in human disease loci. In the second, we investigated the expression profile of a poorly understood cell type, ***enteric glia***, using a combination of CellMapper and RNA-Sequencing. As these examples are independent of one another, I will describe my specific role and contributions to these projects at the beginning of each subsection.

5.1 Prioritizing candidate genes in human disease loci

Genome-wide association studies (GWAS) have linked numerous human genetic variants, such as single nucleotide polymorphisms (SNPs), to different traits and diseases. Although each associated variant implicates a genomic region that can include as many as ten or more genes, only one is typically relevant to disease pathogenesis[89]. There are a

limited set of approaches available to identify which gene(s) surrounding a variant is most likely to contribute to disease, posing a major bottleneck in translating GWAS results into mechanistic insight. As many human diseases are caused by defects in specific tissues or cell types, one fruitful approach has been to identify genes in disease loci that are selectively expressed in the tissue(s) or cell type(s) most relevant to disease[90–92]. CellMapper offers a powerful tool for this type of analysis, because it can be applied to every cell type that is relevant to disease rather than just the cell types that are represented in available large-scale expression datasets. In this first example, we applied CellMapper predictions to prioritize GWAS candidate genes linked to red blood cell[93] and platelet[94] phenotypes, or to inflamatory bowel disease[92] (IBD). This analysis highlighted many candidate genes that were missed by previous approaches, and we experimentally confirmed cell type-selective expression for three of these. I was involved in nearly every stage of this subproject, including performing all analyses and designing the experiments. For the experimental validation, we collaborated with several labs who provided pure samples of the cell types, and then I performed the qRT-PCR.

As an initial case study, we applied CellMapper to prioritize genes from two recent GWAS meta-analyses of erythrocyte[93] and platelet[94] phenotypes, two examples where high quality GWAS data are available and the relevant cell type is unambiguous. Providing initial evidence that CellMapper might be used to highlight genes from these studies, CellMapper predictions for erythrocytes and platelets were >10 fold enriched within 10 kb of SNPs associated with red blood cell and platelet phenotypes, respectively ($p = 2.1 \times 10^{-10}$, $p = 2.3 \times 10^{-5}$; Fisher's exact test). We searched the GWAS loci for erythrocyte and platelet genes, and found 47 candidates predicted to be selectively expressed in the

relevant cell type. One gene that stood out was *TRIM58* because it is in a locus associated with both erythrocyte and platelet cell number (**Fig 5.1A**) and predicted to be selectively expressed in both cell types with high confidence (FDR < 10^{-15}). We measured *TRIM58* expression across hematopoietic cells by qRT-PCR, and found that it was expressed exclusively in erythrocytes, platelets, and their common progenitors (**Fig 5.1B**), validating our predictions and implicating a role for *TRIM58* in the developmental program for platelets and erythrocytes. A functional role for *TRIM58* in erythrocyte development was later confirmed after our analysis[95].

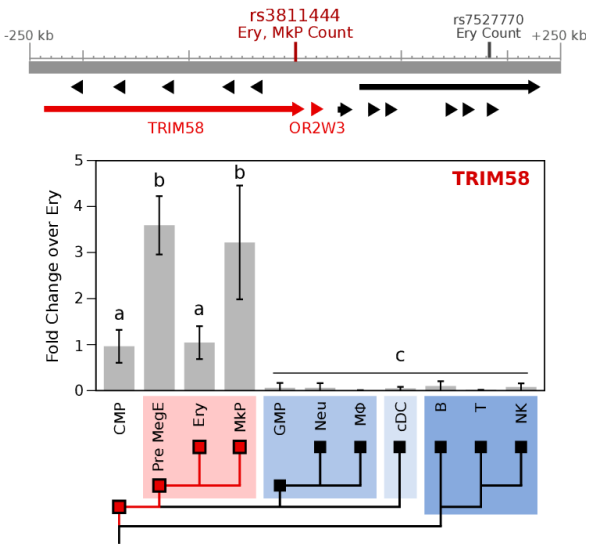


Figure 5.1: Using CellMapper to prioritize GWAS disease genes, part 1. (A) The genetic locus surrounding sentinel SNP rs381144, associated with erythrocyte (Ery) and platelet (MkP) cell number. Genes predicted for expression in erythrocytes and platelets are displayed in red. (B) TRIM58 expression in primary mouse hematopoietic cells as measured by qRT-PCR. MPP, Multi-Potent Progenitor; PreMegE, Pre-Megakaryocyte-Erythrocyte; Ery, Erythrocyte; MkP, Megakaryocyte/Platelet; GMP, Granulocyte-Monocyte Progenitor; Neu, Neutrophil; MΦ, Macrophage; cDC, conventional Dendritic Cell; B, B Cell; T, T Cell; NK, Natural Killer Cell. All bars are mean +/- SD (n = 3 – 7 independent biological replicates), and letters indicate statistically significant differences between groups ($p \leq 0.05$, Tukey's Honest Significant Difference test).

We next applied CellMapper to analyze GWAS results for the chronic inflammatory bowel diseases (IBD), a complex set of diseases involving many cell types, including some that lack gene expression profiles. We focused on the 163 IBD susceptibility loci identified by Jostins, et al.[92], 38 of which lack any candidate gene(s) highlighted by

previous prioritization strategies. Genes predicted by CellMapper to be differentially expressed in T cells, B cells, NK cells, and platelets were more than 5-fold enriched among genes located within 10 kb of IBD SNPs ($p < 0.01$ for all cell types), highlighting the well-known relevance of the three lymphocyte cell types to IBD[96] and supporting the view that platelets also play an active role in disease pathogenesis[97]. We searched IBD loci for genes predicted to be differentially expressed in these four cell types and four others that contribute to IBD[96] – macrophages, simple epithelial cells, goblet cells, and Paneth cells. This analysis highlighted 65 novel candidates and provided additional support for 74 previously implicated genes.

Example candidates highlighted by CellMapper are *C1orf106* and *KIF21B* (**Fig 5.2A**), two genes in the same locus predicted to be enriched in simple epithelial cells and in T and NK cells, respectively. As before, we verified our expression predictions by qRT-PCR, this time using human immune cell types isolated by FACS, cultured endothelial and epithelial cell lines, and primary intestinal epithelial organoids (**Fig 5.2B**). The results

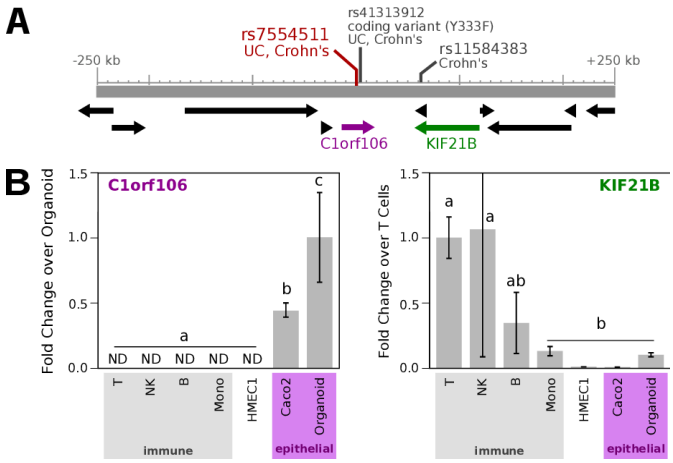


Figure 5.2: Using CellMapper to prioritize GWAS disease genes, part 2. (A) The genetic locus surrounding sentinel SNP rs7554522, associated with inflammatory bowel disease (IBD). Genes colored in purple are predicted for simple epithelial cells, genes colored green predicted for T and NK cells. (B) *C1orf106* and *KIF21B* expression in human primary cells and cell lines. Mono, monocyte; HMEC1, endothelial cell line; Caco2, colon epithelial cell line; Organoid, primary epithelial organoid from small intestine biopsy. All bars are mean +/- SD (n = 3 – 7 independent biological replicates), and letters indicate statistically significant differences between groups ($p \leq 0.05$, Tukey's Honest Significant Difference test).

confirmed epithelial expression of *C1orf106*, and T and NK cell expression of *KIF21B*. This example illustrates one benefit of CellMapper as a prioritization strategy for GWAS: CellMapper can be used to not only prioritize candidate genes, but also to suggest which cell type(s) might be affected for each candidate. *C1orf106*, the gene we discovered to be epithelia-specific, is particularly interesting as an IBD candidate because rare coding variants in this gene have been associated with an increased risk for IBD[98].

To assess whether CellMapper could also be used to prioritize candidates for other diseases, we comprehensively searched for enrichment of disease candidate genes among our top predictions for each of the 30 cell types. We considered genes linked to human genetic disorders in Online Mendelian Inheritance in Man[99] (OMIM) and genes in disease susceptibility loci identified by GWAS[100]. Both OMIM genes and GWAS candidates were significantly enriched in the top 200 predictions across all CellMap cell types ($p = 7.4 \times 10^{-37}$ and 3.3×10^{-30}, respectively). Furthermore, we frequently found that genes linked to individual diseases were enriched in the top predictions for specific cell types (**Fig 5.3A,B**), and these disease-cell type associations primarily highlighted cell types with an established role in disease pathology. These results demonstrate the potential of CellMapper to prioritize genes for many other human diseases.

Figure 5.3: Enrichment of genes linked to (A) human genetic disorders (OMIM) or (B) human GWAS phenotypes (NHGRI) within the top 200 genes prediction for a given cell type. All cell type-disease enrichments that reached statistical significance are shown.

5.2 Transcriptional identity of enteric glia

Glia are the supportive cells of the nervous system. In the enteric (intestinal) nervous system, glia outnumber neurons several fold and are essential for normal intestinal function[101], yet the physiological role of this cell type remains poorly understood. I contributed to a project in collaboration with Drs. Meena Rao and Gabrial Corfas to study gene expression in enteric glia using a combination of RNA-Seq and CellMapper. The goal was to identify new cell type-specific markers for enteric glia and to compare enteric glia to other neural cell types based on expression similarity. Enteric glia have previously been thought to be most similar to astrocytes[102], due to their similar morphology and the fact that they express the astrocyte marker *Gfap*, and to Schwann cells[102], due to their similar developmental origins. However, we found that enteric glia also express many markers of myelinating oligodendrocytes, and bear as much global expression similarity to oligodendrocytes as they do to astrocytes or Schwann cells. These results reveal that enteric glia cannot be considered analogous to either astrocytes or Schwann cells, but are rather transcriptionally distinct from all other neural cell types – with some overlap in expression with other glia, but also many differences. We also identified many new genes enriched in enteric glia, and some of these may be useful as cell type markers. My role in this project was to analyze the RNA-Seq data, plan and conduct the expression comparison between neural cell types, and provide other bioinformatics support. This section was co-written with Dr. Rao, and the project has been previously published[103].

5.2.1 Transcriptional profiling of enteric glia

Genome-wide expression data were not previously available for enteric glia; to gain insight into the functions of glia in the gastrointestinal tract, we determined the transcriptional profile of *Plp1*+ enteric glia by RNA-Seq. GFP+ cells were isolated by FACS from the ileum and colon of Tg(Plp1-GFP) reporter mice – which express GFP on the *Plp1* promoter. Then total RNA abundance was measured by Illumina deep sequencing, and differential expression assessed with Cuffdiff[104]. For the differential expression analysis, GFP+ samples from colon and ileum were treated as biological replicates and compared to GFP- samples. This represents a conservative strategy that should highlight genes selectively expressed in enteric glia in both colon and ileum. To validate our RNA-Seq results, we first checked the measured expression of established marker genes for several intestinal cell types. The enteric glial genes *S100b*, *Gfap*, *Plp1*, and *Sox10* were strongly enriched in GFP+ samples, while markers of epithelial cells, smooth muscle, and endothelia were all depleted (**Fig 5.4**). A subset of neuronal genes that are not expressed by enteric glia were also somewhat enriched in the GFP+ sample (**Fig 5.4**, right), suggesting that this sample contained some neuronal contamination. Regardless, glial genes were enriched to levels at least 20-fold higher than that of neuronal genes in the GFP+ sample; therefore, the presence of some neurons is unlikely to confound further analysis, which focuses on the most differentially expressed genes.

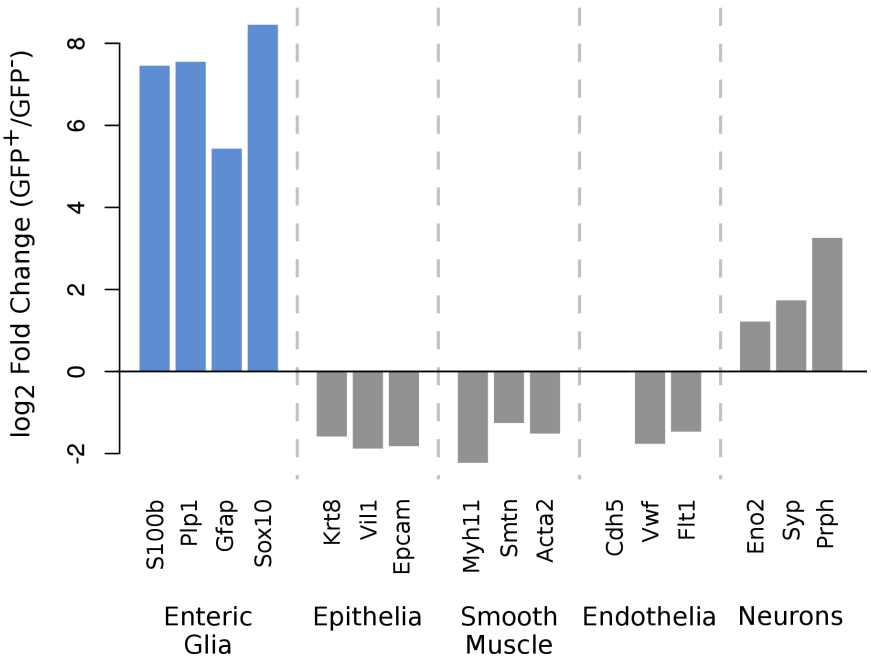


Figure 5.4: RNA from *Plp1*+ cells isolated from postnatal intestine is enriched for enteric glial marker genes. Relative difference in expression of cell-type specific genes in RNA-Seq data from GFP+ versus GFP- samples. To avoid numerical instability (dividing by 0) in cases where a gene was very weakly expressed in one sample, a pseudocount of 0.01 was added to all measured FPKM prior to calculating fold enrichment. This will, in general, decrease the estimated fold change, providing a more conservative value.

5.2.2 Enteric glia are transcriptionally distinct from other types of glia

Enteric glia share the same developmental origins as Schwann cells, but have traditionally been considered to be analogous to astrocytes based on their morphology and expression of *Gfap*[102]. To explore the transcriptional similarity between enteric glia and other types of glia in the nervous system, we compared our data to previous microarray[57,105] and RNA-Seq[58] studies of gene expression in murine astrocytes, neurons, oligodendrocytes,

microglia and Schwann cells. In a second complementary approach, we used CellMapper to predict genes expressed in enteric glia and other neural cell types using human microarray data from ArrayExpress[79,106] and the Allen Brain Atlas[82]. In total, these two approaches provide independent gene expression measurements for each cell type, allowing us to compare our results to several studies and determine which cell type-similarities are robust across methods.

There are several technical challenges when comparing gene expression between studies: different cell isolation protocols, RNA purification methods, and expression technologies, all leading to “lab effects” that can substantially alter an overall expression profile[107]. To overcome these uncertainties, we designed a strategy to distinguish true biological differences between cell types while mitigating technical variability. First, we compared genes upregulated in a given cell type (relative expression) rather than raw expression levels. Second, we assessed similarity of gene expression between each experiment by calculating a “similarity score”[108], which emphasizes genes that are most differentially expressed in an experiment over genes that do not change substantially and contribute to noise. To test this approach, we calculated similarity scores between every pair of studies and estimated how much of the variability in these scores could be explained by the combination of cell types being compared (e.g. Neuron-Neuron, Neuron-Enteric Glia). Cell type combination explained ~80% of the variance in similarity scores (η^2 = 83%; $p < 10^{-4}$ by permutation test), indicating that the similarity score strategy can highlight true biological differences between cell types.

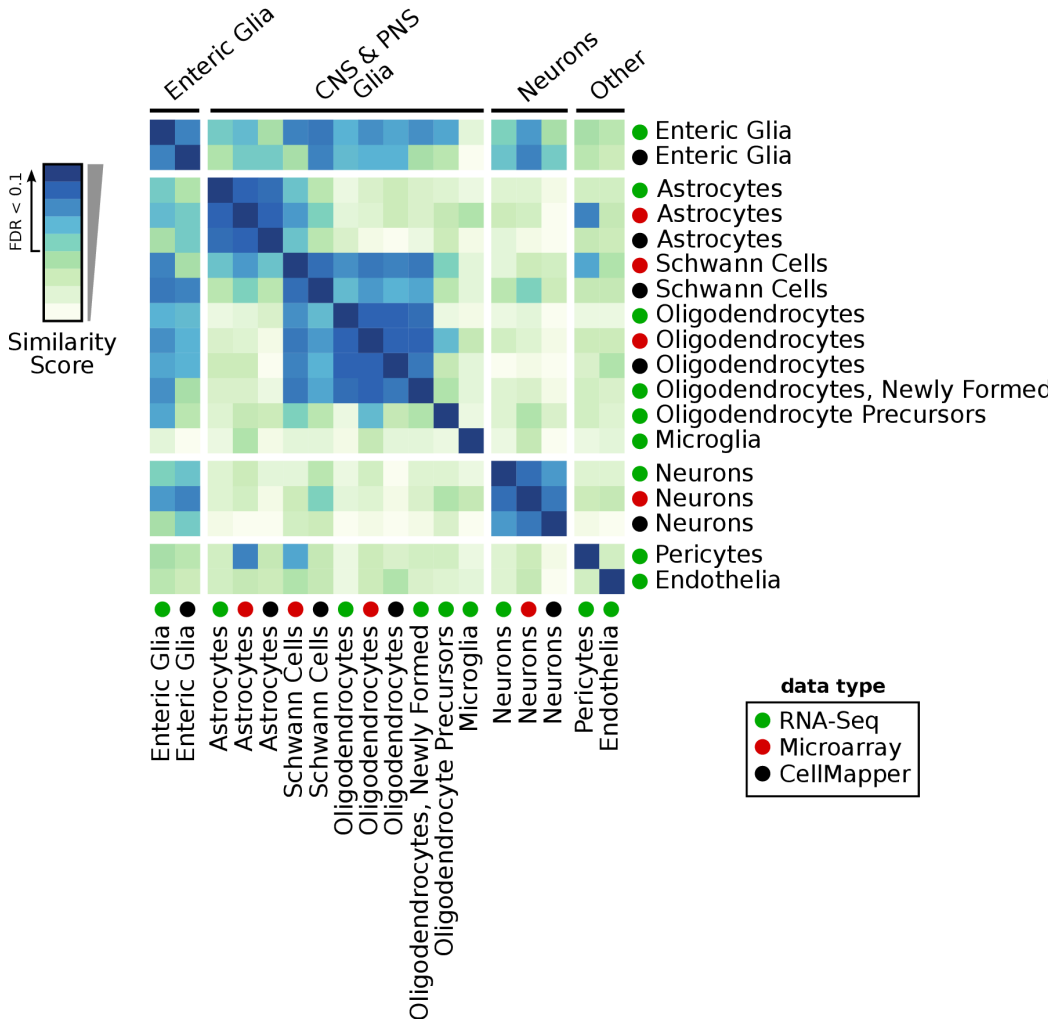


Figure 5.5: Global comparisons of enteric glial transcriptional profiles to that of other neural cell types, part 1: Heatmap illustrating global similarity scores between gene expression in glial and non-glial cell types. (B) Gene set enrichment analysis (GSEA) showing the rank of markers from neurons (N), oligodendrocytes (O), and astrocytes (A) within the transcriptomes of CNS cell types (top) versus enteric glia (bottom).

Similarity scores between every combination of samples are illustrated as a heatmap in **Figure 5.5**. This plot shows the consistency across methods aimed at measuring expression for the same cell type. For example, at a global level, gene expression in astrocytes is highly similar whether measured by RNA-Seq, microarray or CellMapper. Previously known similarities and differences between cell types are also

evident. For instance, oligodendrocytes, astrocytes and neurons have very little gene expression similarity with each other, and consequently similarity scores between these cell types are very low. Also as expected, Schwann cells show strong similarity with oligodendrocytes, the other myelinating cell type, but only weak similarity to astrocytes and no similarity to neurons. Enteric glia display the greatest similarity to Schwann cells, and then to oligodendrocytes. They also exhibit some similarity to astrocytes, but it is limited and no greater than the similarity between Schwann cells and astrocytes. Enteric glia share no significant similarity with the mesoderm-derived microglia, pericytes or endothelia. Of note, enteric glia exhibit significant similarity scores with neurons, unlike any of the other types of glia. It is unclear if this similarity is due to neuronal contamination of the enteric glial data set, or to shared expression of some biological pathways. The consistency of this observation across the RNA-Seq and CellMapper data supports the possibility of shared pathways.

The global comparisons of transcriptional similarity suggest that enteric glia express many of the same genes as several different types of glia. To investigate this suggestion further, we examined our RNA-Seq and CellMapper predictions for expression of an established set of astrocyte, oligodendrocyte, and neuron markers[57], and asked whether each marker set was enriched in enteric glia according to gene set enrichment analysis[109] (GSEA). To validate this approach, we first assessed the rank of these markers within RNA-Seq data sets for each of the CNS cell types[58]. We found that neuronal genes were enriched only in neurons, astrocytic genes in astrocytes, and oligodendrocyte genes in oligodendrocytes (**Fig 5.6**, top), confirming that the “established markers” chosen are consistent across studies and represent an accurate set of cell type-specific markers

within the CNS.

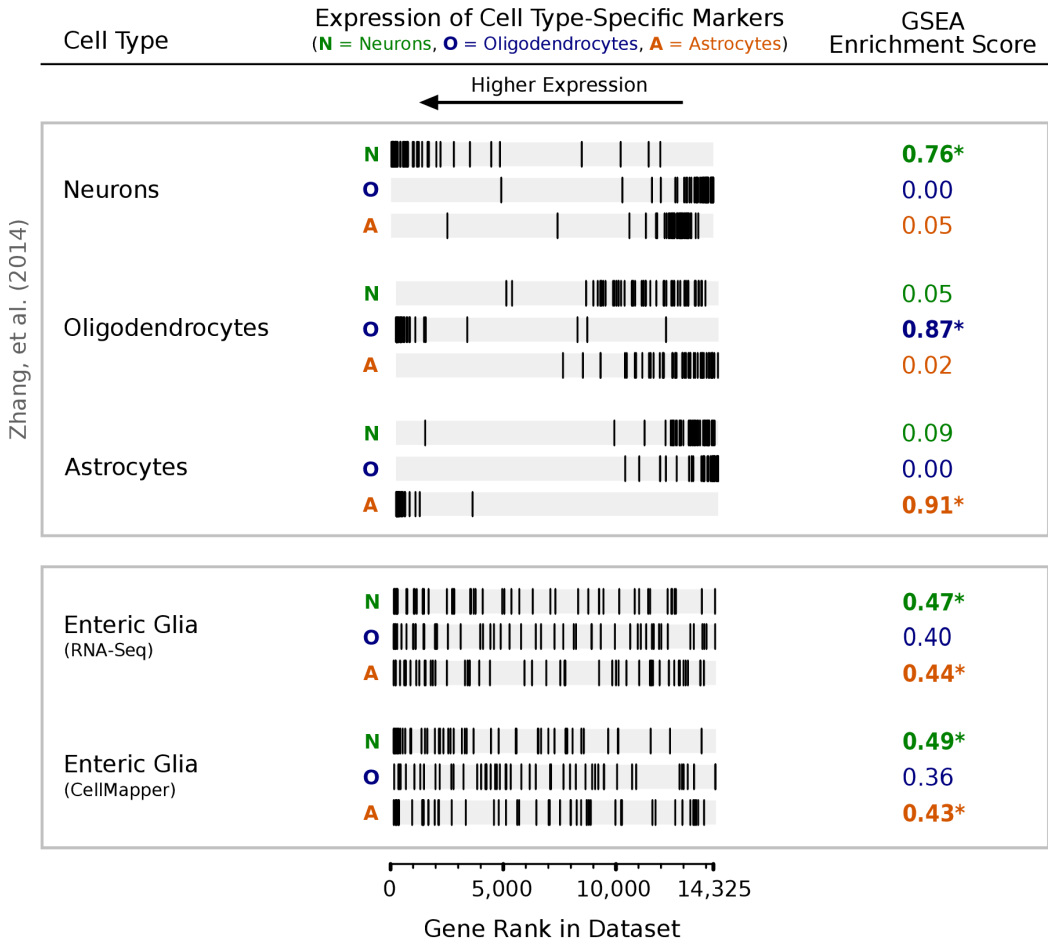
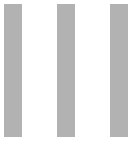


Figure 5.5: Global comparisons of enteric glial transcriptional profiles to that of other neural cell types, part 2: Gene set enrichment analysis (GSEA) showing the rank of established markers[57] from neurons (N), oligodendrocytes (O), and astrocytes (A) within the transcriptomes of CNS cell types (top) versus enteric glia (bottom).

When we carried out GSEA using these markers in the enteric glia data set, the results were striking. Markers for all 3 CNS cell types were enriched, with some markers being strongly expressed in enteric glia (ranking among the top enriched enteric glial genes), while other markers were undetected (**Fig 5.6**, bottom). This result was equivalent

whether we analyzed the enteric glial expression profile predicted by CellMapper or the RNA-Seq data, showing that the finding was consistent across methods. Certain subsets of markers of each CNS cell type were strongly enriched in our GFP+ sample. For instance, a subset of genes important in myelinating glia including *Sox10*, *Plp1*, *Mbp*, and *Mpz*, were all strongly enriched while others, such as *Mog*, *Mobp*, and *Mag* were not detected. Astrocytic genes such as *Gfap*, *Entpd2*, and *Dio2* were all highly enriched in enteric glia while other widely used markers of astrocytes, such as *Aldh1l1* and the glutamate transporter *Slc1a3*, were not expressed. Taken together, the GSEA and global analyses of transcriptional similarity show that enteric glia are a unique class of glia, without direct analogy to any other type of glial cell examined.



Appendix

A

Details of Normalization and Analysis for RNAi Screen

To analyze our data, we developed a strategy that assesses esiRNA activity at the same time as normalization, estimating the consistent effect among all esiRNA replicates which cannot be explained by known systematic sources of variation. Let $\boldsymbol{X_{e\phi i}}$ be the log-transformed luciferase intensity of well ***i***, treated with esiRNA ***e***, and measured at location $\boldsymbol{\phi}$ (ϕ = {b,p,r,c} for batch ***b***, plate ***p***, row ***r***, and column ***c***). We fit the measured intensities, $X_{e\phi i}$, to a mixed effects linear model with fixed esiRNA effects $\boldsymbol{\mu_e}$, random systematic effects $\boldsymbol{S_\phi}$, and residual error $\boldsymbol{\varepsilon_i}$:

$$X_{e\phi i} = \mu_e + S_\phi + \varepsilon_i$$

where systematic effects are the sum of batch effects $\boldsymbol{B_b}$, plate effects $\boldsymbol{P_p}$, row effects $\boldsymbol{R_{pr}}$, and column effects $\boldsymbol{C_{pc}}$:

$$S_\phi = B_b + P_p + R_{pr} + C_{pc}$$

This model was fit separately for each direction of transport. Normalized signal intensities, which have been corrected for systematic errors ($X_{e\phi i} - S_\phi$), are plotted in **Figure A.1B**.

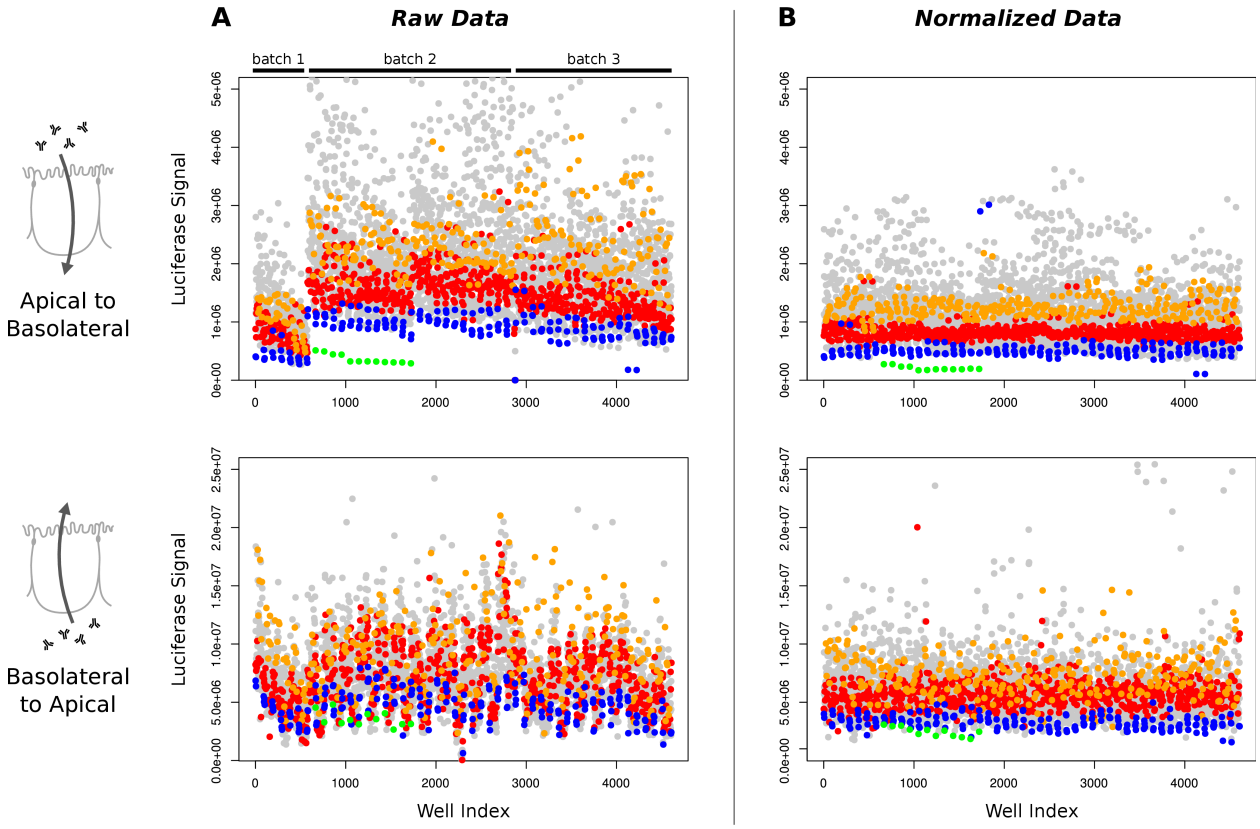


Figure A.1: Raw (A) and normalized (B) results from the screen. Each dot corresponds to a single luciferase measurement; there are 6 measurements for each esiRNA (3 biological x 2 technical replicates). Blue, Pos (GFP - Eupheria); Green, GFP - Lencer; Red, Neg1 (mock); Orange, Neg2 (Luc); Gray, other.

To assess statistical significance of each esiRNA effect μ_e, we used a permutation test. A single mock transfection well on the library plate was labeled as a gene named “MOCK”, and then the mixed linear model fit was repeated, obtaining an estimate of the “esiRNA” effect for a known negative control, $\boldsymbol{\mu_{MOCK}}$. This was then repeated for the remaining mock transfection wells (112 in total), providing an empirical null distribution under the mock transfection condition. The null distribution, μ_{MOCK}, fits a normal distribution for both directs of transport ($p = .82, .40$ for basolateral to apical and apical to basolateral transport, Shapiro-Wilk test). As can be seen in **Figure A.2**, the actual distribution of esiRNA effects is spread out relative to the null distribution.

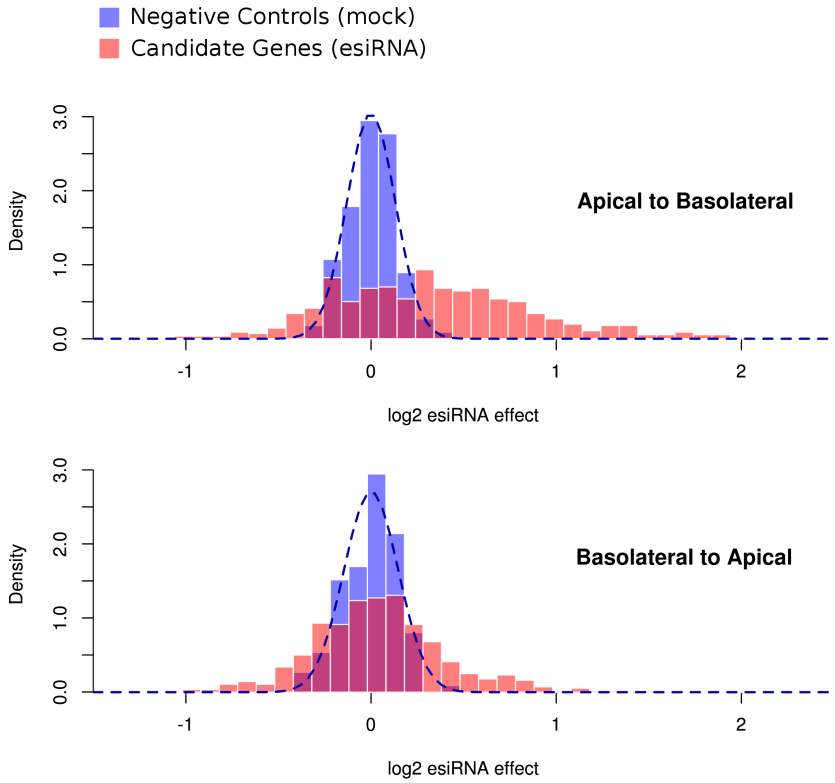


Figure A.1: Histogram of the $\log_2$ esiRNA effect, μ_e, is shown in red. Histogram of empirical null distribution, μ_{MOCK}, is shown in blue, with the best fit normal distribution plotted as a dotted blue line.

A Z-score was then calculated for each gene and direction of transport by dividing the estimated esiRNA effect, μ_e, by the standard deviation of the empirical null distribution (i.e. the standard deviation of μ_{MOCK}). Z-scores for each direction were then pooled using Stouffer's method:

$$Z_e = \frac{Z_e^{A2B} + Z_e^{B2A}}{\sqrt{2}}$$

resulting in a single overall Z-score for each esiRNA in the screen.

B

Comparing CellMapper to Other Approaches

This section provides a performance comparison between CellMapper and several other methods that can be applied to identify genes enriched in different cell types, including the unsupervised hierarchical clustering algorithm weighted gene co-expression network analysis (WGCNA)[110], several complete deconvolution algorithms[55], and Pearson's correlation. Although these algorithms can all be used to identify genes expressed selectively in specific cell types, most were not designed for this purpose: WGCNA was designed to explore large patterns in gene expression data, and complete deconvolution algorithms were designed to separate expression changes due to variation in cell proportions and changes due to expression differences within the individual cell types. CellMapper cannot address these alternative problems, and the performance evaluations below only demonstrate the superiority of CellMapper for the specific, but important challenge of identifying cell type-enriched genes. A comparison between CellMapper and the machine-learning algorithm, *in silico* nano-dissection[65], is provided in Chapter 4.2.3.

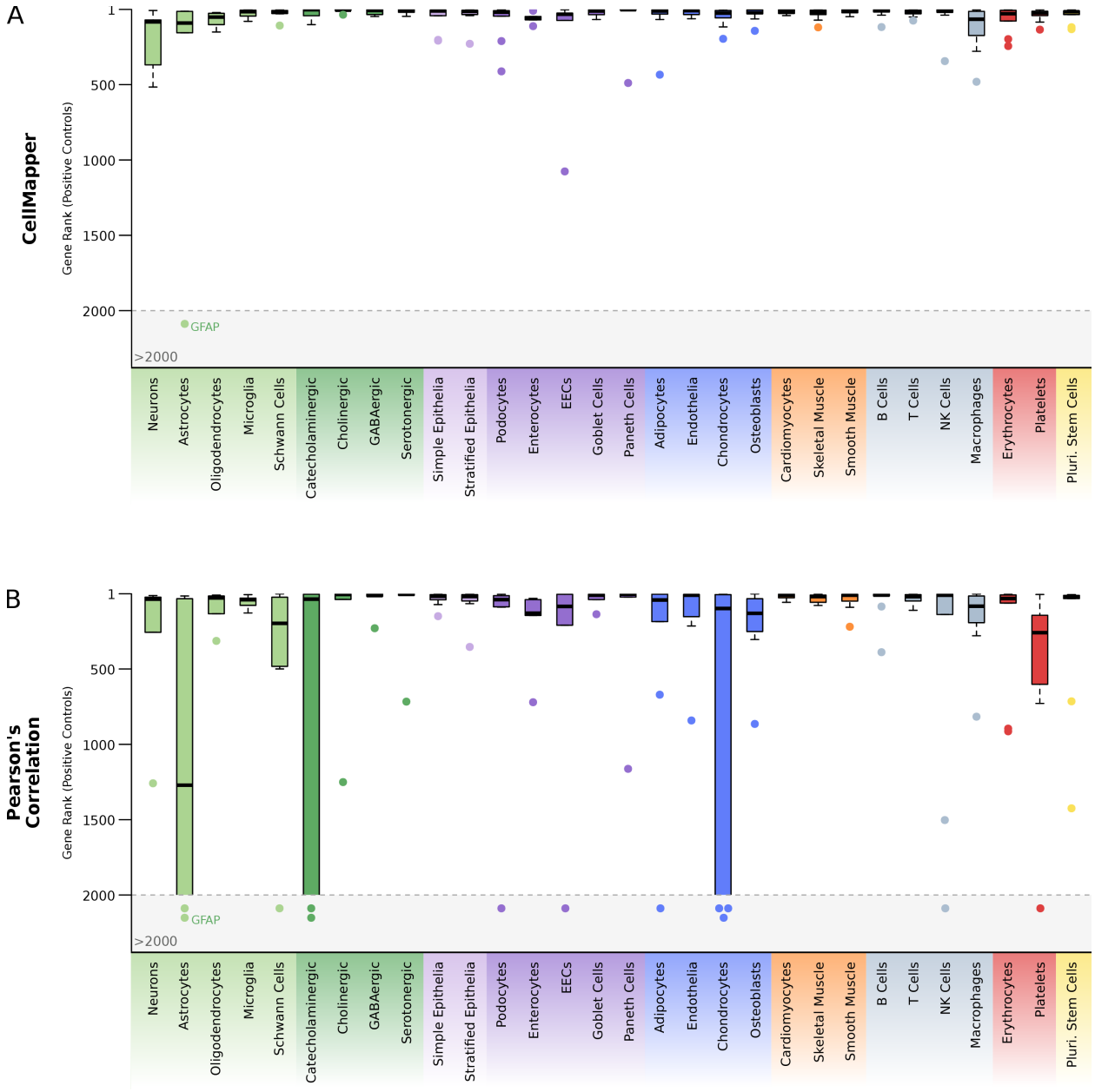


Figure B.1: Comparison to Pearson's correlation. All searches described in **Table 4.2** were repeated with Pearson's correlation. Tukey boxplots display the rank of 5-10 literature curated markers (positive controls) for each of the 30 cell types. While CellMapper was accurate for every cell type (A), Pearson's correlation performed poorly for about a third of cell types (B). **Figure 4.6** shows the same data as panel A in log scale, this figure is in linear scale to more clearly demonstrate the poor performance of Pearson's correlation for many cell types. *GFAP*, the only positive control gene ranked > 2000 by CellMapper, was also ranked > 2000 by Pearson's correlation. EECs, enteroendocrine cells.

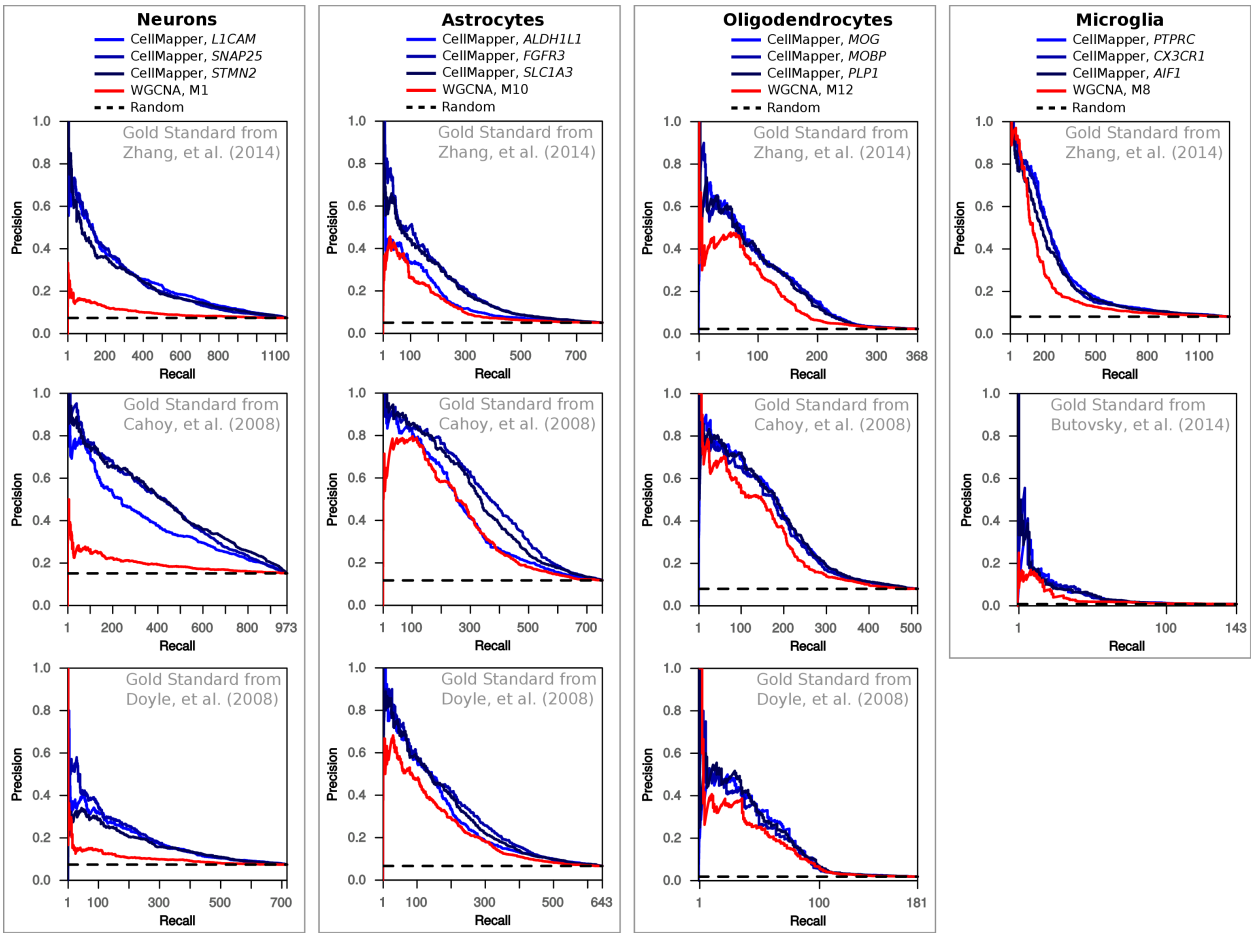


Figure B.2: Comparing CellMapper to unsupervised clustering, part 1. Unsupervised clustering methods, such as Weighted Gene Coexpression Network Analysis (WGCNA), group genes into modules based on their similarity of expression. These methods provide a powerful means to identify major patterns of coexpression in microarray data. WGCNA has been shown to uncover modules in whole brain tissue data that correspond to genes expressed in specific cell types[82,110], and so we decided to compare the accuracy of this unsupervised approach to CellMapper. Precision recall plots comparing WGCNA to CellMapper, based on the recovery of experimentally defined ("gold standard") cell type-enriched genes[56–58,83]. CellMapper produces a small, but reproducible performance increase for astrocytes, oligodendrocytes, and microglia, and a large performance increase for neurons. The best performing WGCNA module is plotted for each cell type.

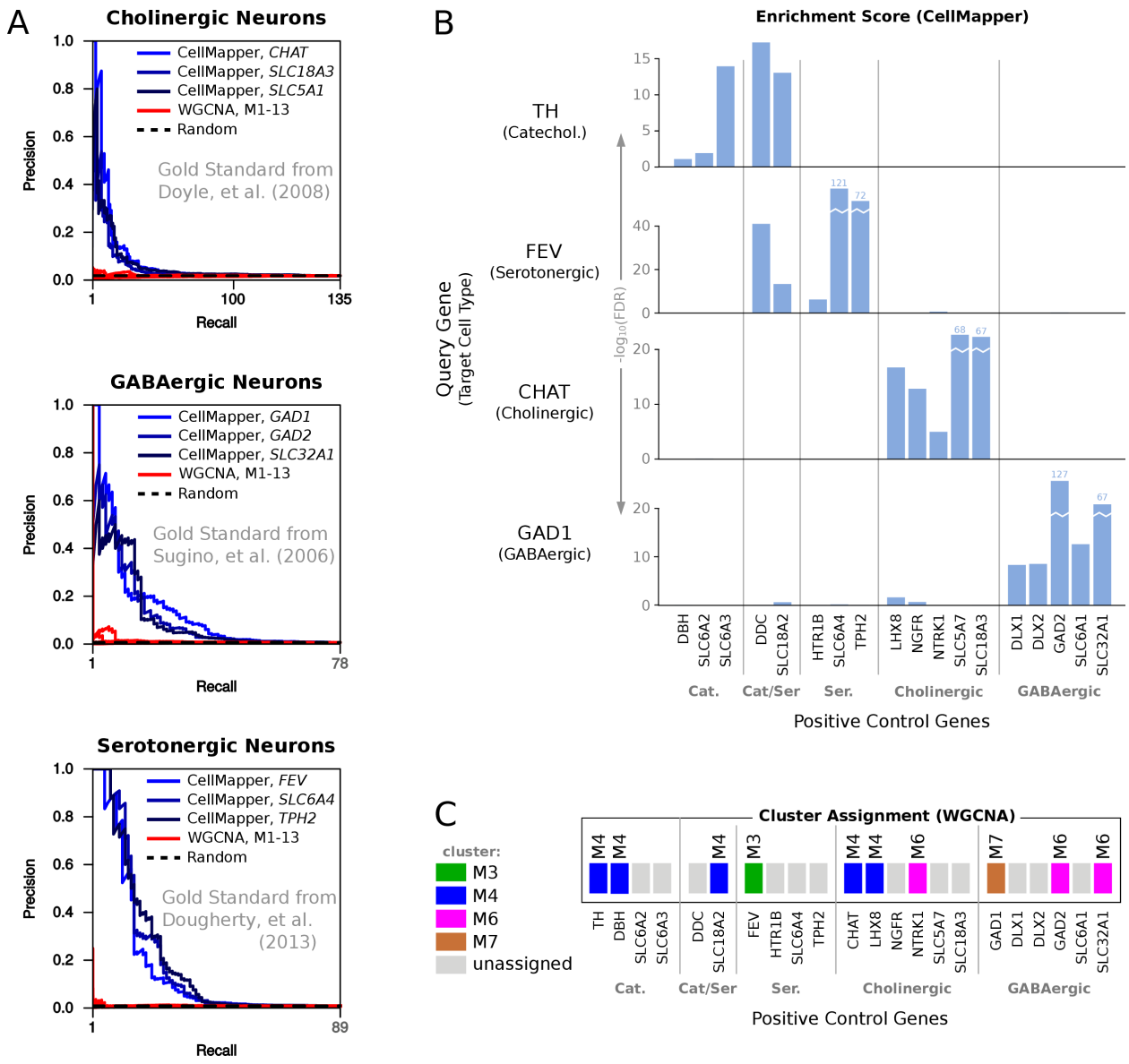


Figure B.3: Comparing CellMapper to unsupervised clustering, part 2. While unsupervised clustering methods, such as WGCNA, can uncover modules in whole brain tissue data that correspond to genes expressed in the major cell classes, they offer no ability to target specific cell types. (A) Comparing the ability of CellMapper and WGCNA to recover experimentally-defined neuron subtype genes[56,71,72]. Precision recall curves confirm that CellMapper recovers genes in expressed neuron subtypes using multiple query genes, while WGCNA does not recover any modules related to these cell types. Precision recall curves are plotted for all 13 WGCNA modules, demonstrating that no WGCNA modules recover neuron subtype genes better than expected by random chance. (B) The correct identification of literature curated (positive control) cell-specific markers confirm that CellMapper accurately distinguished genes between the four neuron subtypes. The positive control genes *DDC* and *SLC18A2*, which are

Figure B.3 (cont.): expressed selectively in both catecholaminergic and serotonergic neurons, were correctly associated with both cell types. (C) In contrast, WGCNA did not identify any gene modules related to these four neuron subtypes, as shown by the lack of correspondence between module assignment and known association with these cell types. Cat, catecholaminergic neurons; Ser, serotonergic neurons; Cholinergic, cholinergic neurons; GABAergic, GABAergic neurons.

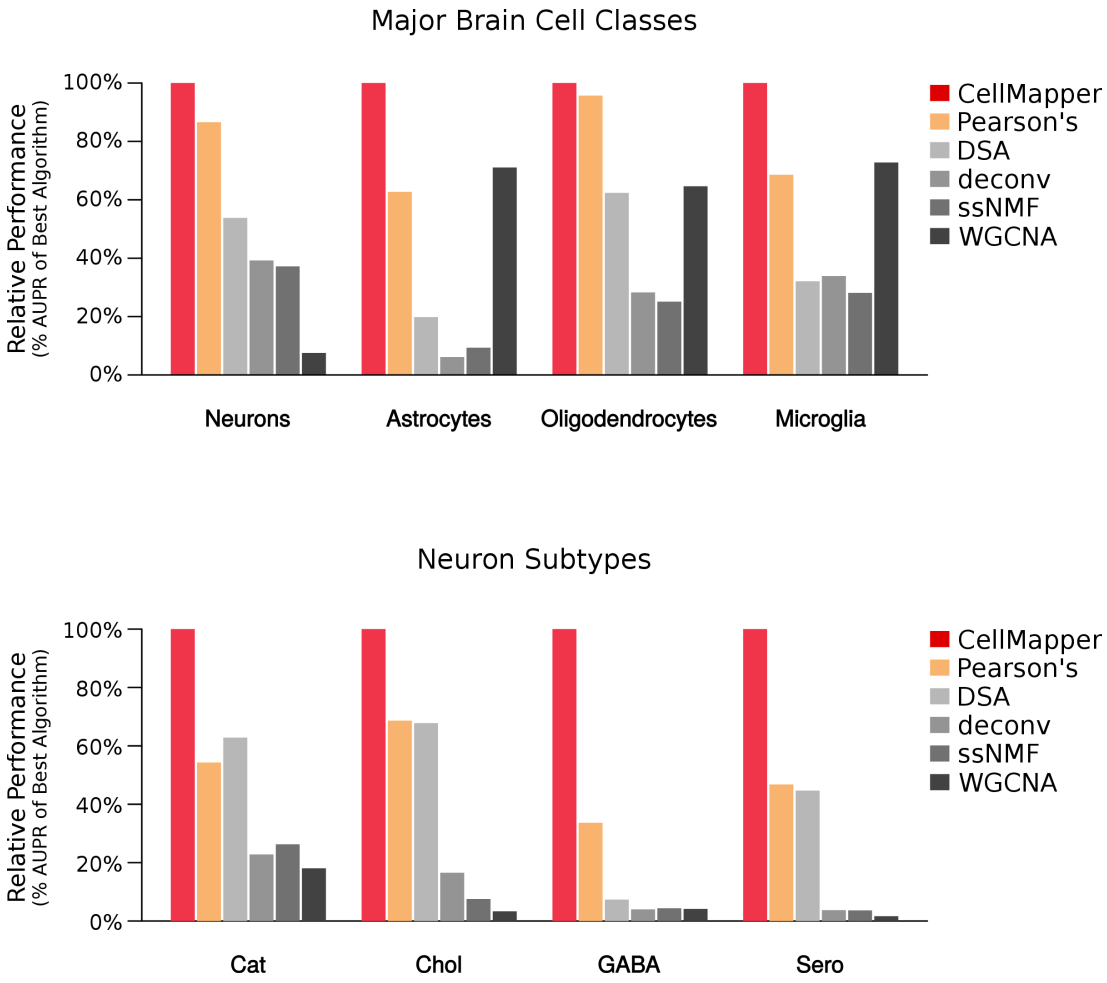


Figure B.4: Comparing CellMapper to complete deconvolution algorithms. All complete deconvolution algorithms from the CellMix R package[113] were tested, which includes the Digital Sorting Algorithm[70] (DSA), deconv[111], and semi-supervised Non-negative Matrix Factorization[112] (ssNMF), along with Pearson's correlation and WGCNA[82]. For this evaluation, we applied each method to predict genes selectively expressed in the four major brain cell classes from **Figure B.2** (top) and four neuron subtypes from **Figure B.3** (bottom) using microarray data from the Allen Brain Atlas[82]. Then we determined how accurately each method returned an experimentally defined set of genes for each cell type[56,58,71,72,114], as quantified by AUPR. "Relative Performance" is calculated from the AUPR, linearly scaled so that the best performing algorithm for each cell type is given a value of 100%, and the AUPR expected by random chance is 0%. Many cell type deconvolution algorithms were originally validated using the major brain cell classes, and are expected to perform well for these cases. The neuron subtypes, in contrast, have not been successfully analyzed by other methods.

References

1. Rodriguez-Boulan, E., Kreitzer, G. & Müsch, A. Organization of vesicular trafficking in epithelia. *Nat. Rev. Mol. Cell Biol.* **6,** 233–47 (2005).

2. Tuma, P. L. & Hubbard, A. L. Transcytosis: crossing cellular barriers. *Physiol. Rev.* **83,** 871–932 (2003).

3. Bomsel, M. Transcytosis of infectious human immunodeficiency virus across a tight human epithelial cell line barrier. *Nat. Med.* **3,** 42–47 (1997).

4. Couesnon, A., Pereira, Y. & Popoff, M. R. Receptor-mediated transcytosis of botulinum neurotoxin A through intestinal cell monolayers. *Cell. Microbiol.* **10,** 375–387 (2008).

5. Roopenian, D. C. & Akilesh, S. FcRn: the neonatal Fc receptor comes of age. *Nat. Rev. Immunol.* **7,** 715–25 (2007).

6. Tzaban, S. *et al.* The recycling and transcytotic pathways for IgG transport by FcRn are distinct and display an inherent polarity. *J. Cell Biol.* **185,** 673–684 (2009).

7. Kuo, T. T. *et al.* Neonatal Fc receptor: From immunity to therapeutics. *J. Clin. Immunol.* **30,** 777–789 (2010).

8. Yeung, Y. A. *et al.* Engineering human IgG1 affinity to human neonatal Fc receptor: impact of affinity improvement on pharmacokinetics in primates. *J. Immunol.* **182,** 7663–7671 (2009).

9. Zalevsky, J. *et al.* Enhanced antibody half-life improves in vivo activity. *Nat. Biotechnol.* **28,** 157–159 (2010).

10. Leung, S. M., Ruiz, W. G. & Apodaca, G. Sorting of membrane and fluid at the apical pole of polarized Madin-Darby canine kidney cells. *Mol. Biol. Cell* **11,** 2131–2150 (2000).

11. Parton, R. G., Prydz, K., Bomsel, M., Simons, K. & Griffiths, G. Meeting of the apical and basolateral endocytic pathways of the Madin-Darby canine kidney cell in late endosomes. *J. Cell Biol.* **109,** 3259–3272 (1989).

12. Bomsel, M., Parton, R., Kuznetsov, S. a, Schroer, T. a & Gruenberg, J. Microtubule- and motor-dependent fusion in vitro between apical and basolateral endocytic vesicles from MDCK cells. *Cell* **62,** 719–31 (1990).

13. Hughson, E. J. & Hopkins, C. R. Endocytic pathways in polarized Caco-2 cells: identification of an endosomal compartment accessible from both apical and basolateral surfaces. *J. Cell Biol.* **110,** 337–48 (1990).

14. Hoekstra, D., Tyteca, D. & van IJzendoorn, S. C. D. The subapical compartment: a traffic center in membrane polarity development. *J. Cell Sci.* **117,** 2183–2192 (2004).

15. Brown, P. S. *et al.* Definition of distinct compartments in polarized Madin-Darby canine kidney (MDCK) cells for membrane-volume sorting, polarized sorting and apical recycling. *Traffic* **1,** 124–40 (2000).

16. Wang, E. *et al.* Apical and basolateral endocytic pathways of MDCK cells meet in acidic common endosomes distinct from a nearly-neutral apical recycling endosome. *Traffic* **1,** 480–93 (2000).

17. Thompson, A. *et al.* Recycling Endosomes of Polarized Epithelial Cells Actively Sort Apical and Basolateral Cargos into Separate. **18,** 2687–2697 (2007).

18. Zerial, M. & McBride, H. Rab proteins as membrane organizers. *Nat. Rev. Mol. Cell Biol.* **2,** 107–17 (2001).

19. Di Paolo, G. & De Camilli, P. Phosphoinositides in cell regulation and membrane dynamics. *Nature* **443,** 651–657 (2006).

20. Rink, J., Ghigo, E., Kalaidzidis, Y. & Zerial, M. Rab conversion as a mechanism of progression from early to late endosomes. *Cell* **122,** 735–749 (2005).

21. Martin-Belmonte, F. *et al.* PTEN-Mediated Apical Segregation of Phosphoinositides Controls Epithelial Morphogenesis through Cdc42. *Cell* **128,** 383–397 (2007).

22. Gassama-Diagne, A. *et al.* Phosphatidylinositol-3,4,5-trisphosphate regulates the formation of the basolateral plasma membrane in epithelial cells. *Nat. Cell Biol.* **8,** 963–970 (2006).

23. Golachowska, M. R., Hoekstra, D. & van IJzendoorn, S. C. D. Recycling endosomes in apical plasma membrane domain formation and epithelial cell polarity. *Trends Cell Biol.* **20,** 618–26 (2010).

24. Matter, K., Hunziker, W. & Mellman, I. Basolateral sorting of LDL receptor in MDCK cells: the cytoplasmic domain contains two tyrosine-dependent targeting determinants. *Cell* **71,** 741–53 (1992).

25. Deborde, S. *et al.* Clathrin is a key regulator of basolateral polarity. *Nature* **452,** 719–723 (2008).

26. Gan, Y., McGraw, T. E. & Rodriguez-Boulan, E. The epithelial-specific adaptor AP1B mediates post-endocytic recycling to the basolateral membrane. *Nat. Cell Biol.* **4,** 605–609 (2002).

27. Fölsch, H., Ohno, H., Bonifacino, J. S. & Mellman, I. A novel clathrin adaptor complex mediates basolateral targeting in polarized epithelial cells. *Cell* **99,** 189–198 (1999).

28. Cao, X., Surma, M. a & Simons, K. Polarized sorting and trafficking in epithelial cells. *Cell Res.* **22,** 793–805 (2012).

29. Delacour, D. *et al.* Requirement for galectin-3 in apical protein sorting. *Curr. Biol.* **16,** 408–

414 (2006).

30. Simons, K. & Ikonen, E. Functional rafts in cell membranes. *Nature* **387,** 569–72 (1997).

31. Weisz, O. a & Rodriguez-Boulan, E. Apical trafficking in epithelial cells: signals, clusters and motors. *J. Cell Sci.* **122,** 4253–66 (2009).

32. Kittler, R. *et al.* An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division. *Nature* **432,** 1036–1040 (2004).

33. Kittler, R. *et al.* Genome-wide resources of endoribonuclease-prepared short interfering RNAs for specific loss-of-function studies. *Nat. Methods* **4,** 337–344 (2007).

34. Ang, S. F. & Fölsch, H. The role of secretory and endocytic pathways in the maintenance of cell polarity. *Essays Biochem.* **53,** 29–40 (2012).

35. Fölsch, H. Regulation of membrane trafficking in polarized epithelial cells. *Curr. Opin. Cell Biol.* **20,** 208–213 (2008).

36. Grant, B. D. & Donaldson, J. G. Pathways and mechanisms of endocytic recycling. *Nat. Rev. Mol. cell Biol.* **10,** 597–608 (2009).

37. Hsu, V. W. & Prekeris, R. Transport at the recycling endosome. *Curr. Opin. Cell Biol.* **22,** 528–34 (2010).

38. Mostov, K., Su, T. & Beest, M. Polarized epithelial membrane traffic: conservation and plasticity. *Nat. Cell Biol.* **5,** (2003).

39. Perret, E., Lakkaraju, A., Deborde, S., Schreiner, R. & Rodriguez-Boulan, E. Evolving endosomes: how many varieties and why? *Curr. Opin. Cell Biol.* **17,** 423–34 (2005).

40. Cao, Z. *et al.* Use of fluorescence-activated vesicle sorting for isolation of Naked2-associated, basolaterally targeted exocytic vesicles for proteomics analysis. *Mol. Cell. Proteomics* **7,** 1651–67 (2008).

41. Foster, L. J. *et al.* A Mammalian Organelle Map by Protein Correlation Profiling. *Cell* **125,** 187–199 (2006).

42. Lapierre, L. A. *et al.* Characterization of immunoisolated human gastric parietal cells tubulovesicles : identification of regulators of apical recycling. **2733,** (2007).

43. Huh, W.-K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425,** 686–91 (2003).

44. Balklava, Z., Pant, S., Fares, H. & Grant, B. D. Genome-wide analysis identifies a general requirement for polarity proteins in endocytic traffic. *Nat. Cell Biol.* **9,** 1066–73 (2007).

45. Collinet, C. *et al.* Systems survey of endocytosis by multiparametric image analysis. *Nature* **464,** 243–9 (2010).

46. Pinto, F. L. & Lindblad, P. A guide for in-house design of template-switch-based 5’ rapid amplification of cDNA ends systems. *Anal. Biochem.* **397,** 227–232 (2010).

47. Spiess, A.-N. & Ivell, R. A highly efficient method for long-chain cDNA synthesis using trehalose and betaine. *Anal. Biochem.* **301,** 168–174 (2002).

48. Schütze, T. *et al.* A streamlined protocol for emulsion polymerase chain reaction and subsequent purification. *Anal. Biochem.* **410,** 155–157 (2011).

49. Zhang, X. D. *et al.* Integrating experimental and analytic approaches to improve data quality in genome-wide RNAi screens. *J. Biomol. Screen.* **13,** 378–89 (2008).

50. Yu, D. *et al.* Noise reduction in genome-wide perturbation screens using linear mixed-effect models. *Bioinformatics* **27,** 2173–80 (2011).

51. Oztan, A. *et al.* Exocyst Requirement for Endocytic Traffic Directed Toward the Apical and Basolateral Poles of Polarized MDCK Cells. **18,** 3978–3992 (2007).

52. Bryant, D. M. *et al.* A molecular network for de novo generation of the apical surface and lumen. *Nat. Cell Biol.* **12,** 1035–45 (2010).

53. Okaty, B. W., Sugino, K. & Nelson, S. B. A quantitative comparison of cell-type-specific microarray gene expression profiling methods in the mouse brain. *PLoS One* **6,** e16493 (2011).

54. The_FANTOM_Consortium. A promoter-level mammalian expression atlas. *Nature* **507,** 462–70 (2014).

55. Shen-Orr, S. S. & Gaujoux, R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.* **25,** 571–578 (2013).

56. Doyle, J. P. *et al.* Application of a Translational Profiling Approach for the Comparative Analysis of CNS Cell Types. *Cell* 749–762 (2008). doi:10.1016/j.cell.2008.10.029

57. Cahoy, J. D. *et al.* A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J. Neurosci.* **28,** 264–78 (2008).

58. Zhang, Y. *et al.* An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. *J. Neurosci.* **34,** 11929–47 (2014).

59. Rossner, M. J. *et al.* Global transcriptome analysis of genetically identified neurons in the adult cortex. *J. Neurosci.* **26,** 9956–9966 (2006).

60. Jaitin, D. a. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science (80-.).* **343,** 776–779 (2014).

61. Budakian, R. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. **25,** 279–284 (2014).

62. Saliba, A.-E., Westermann, A. J., Gorski, S. a. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* **42,** 8845–8860 (2014).

63. Nelander, S., Mostad, P. & Lindahl, P. Prediction of cell type-specific gene modules: identification and initial characterization of a core set of smooth muscle-specific genes. *Genome Res.* **13,** 1838–54 (2003).

64. Chikina, M. D., Huttenhower, C., Murphy, C. T. & Troyanskaya, O. G. Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans. *PLoS Comput. Biol.* **5,** e1000417 (2009).

65. Ju, W. *et al.* Defining cell-type specificity at the transcriptional level in human disease. *Genome Res.* **23,** 1862–73 (2013).

66. Teng, S., Yang, J. Y. & Wang, L. Genome-wide prediction and analysis of human tissue-selective genes using microarray expression data. *BMC Med. Genomics* **6 Suppl 1,** S10 (2013).

67. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 1–10 (2015).

68. Shen-Orr, S. S. *et al.* Cell type-specific gene expression differences in complex tissues. *Nat. Methods* **7,** 287–9 (2010).

69. Kuhn, A., Thu, D., Waldvogel, H. J., Faull, R. L. M. & Luthi-carter, R. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat. Methods* **8,** 945–7 (2011).

70. Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M. & Liu, Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* **14,** 89 (2013).

71. Sugino, K. *et al.* Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nat. Neurosci.* **9,** 99–107 (2006).

72. Dougherty, J. D. *et al.* The Disruption of Celf6, a Gene Identified by Translational Profiling of Serotonergic Neurons, Results in Autism-Related Behaviors. *J. Neurosci.* **33,** 2732–2753 (2013).

73. Vickaryous, M. K. & Hall, B. K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev. Camb. Philos. Soc.* **81,** 425–55 (2006).

74. Liu, X., Yu, X., Zack, D. J., Zhu, H. & Qian, J. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics* **9,** 271 (2008).

75. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445,** 168–76 (2007).

76. Owen, A. B., Stuart, J., Mach, K., Villeneuve, A. M. & Kim, S. A gene recommender algorithm to identify coexpressed genes in C. elegans. *Genome Res.* **13,** 1828–37 (2003).

77. Adler, P. *et al.* Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biol.* **10,** R139 (2009).

78. Hibbs, M. a *et al.* Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* **23,** 2692–9 (2007).

79. Lukk, M. *et al.* A global map of human gene expression. *Nat. Biotechnol.* **28,** 322–4 (2010).

80. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 10101–6 (2000).

81. Nielsen, T. O. *et al.* Mechanisms of disease Molecular characterisation of soft tissue tumours : a gene expression study. **359,** (2002).

82. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489,** 391–9 (2012).

83. Butovsky, O. *et al.* Identification of a unique TGF-β-dependent molecular and functional signature in microglia. *Nat. Neurosci.* **17,** 131–43 (2014).

84. Elmore, M. R. P. *et al.* Colony-stimulating factor 1 receptor signaling is necessary for microglia viability, unmasking a microglia progenitor cell in the adult brain. *Neuron* **82,** 380–97 (2014).

85. Budakian, R. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (80-.).* **25,** 279–284 (2014).

86. The Jackson Laboratory Cre Repository. at <http://cre.jax.org/>

87. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29,** 569–574 (2013).

88. Brunskill, E. W., Georgas, K., Rumballe, B., Little, M. H. & Potter, S. S. Defining the molecular character of the developing and adult kidney podocyte. *PLoS One* **6,** 1–12 (2011).

89. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5,** e1000534 (2009).

90. Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47,** (2015).

91. Miller, J. a, Horvath, S. & Geschwind, D. H. Divergence of human and mouse brain

transcriptome highlights Alzheimer disease pathways. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 12698–703 (2010).

92. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–24 (2012).

93. Van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **1,** (2012).

94. Gieger, C. *et al.* New gene functions in megakaryopoiesis and platelet formation. *Nature* **480,** 201–8 (2011).

95. Thom, C. S. *et al.* Trim58 Degrades Dynein and Regulates Terminal Erythropoiesis. *Dev. Cell* **30,** 688–700 (2014).

96. Khor, B., Gardet, A. & Xavier, R. J. Genetics and pathogenesis of inflammatory bowel disease. *Nature* **474,** 307–17 (2011).

97. Danese, S., Motte Cd, C. D. La & Fiocchi, C. Platelets in inflammatory bowel disease: clinical, pathogenic, and therapeutic implications. *Am. J. Gastroenterol.* **99,** 938–45 (2004).

98. Rivas, M. a *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* **43,** 1066–73 (2011).

99. Online Mendelian Inheritance in Man, OMIM. *McKusick-Nathans Inst. Genet. Med. Johns Hopkins Univ. (Baltimore, MD)* at <http://omim.org/>

100. Hindorff, L. *et al.* A Catalog of Published Genome-Wide Association Studies. at <www.genome.gov/gwastudies>

101. Bush, T. G. *et al.* Fulminant jejuno-ileitis following ablation of enteric gila in adult transgenic mice. *Cell* **93,** 189–201 (1998).

102. Gershon, M. D. & Rothman, T. P. Enteric glia. *Glia* **4,** 195–204 (1991).

103. Rao, M. *et al.* Enteric glia express proteolipid protein 1 and are a transcriptionally unique population of glia in the mammalian nervous system. *Glia* n/a–n/a (2015). doi:10.1002/glia.22876

104. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–78 (2012).

105. Buchstaller, J. *et al.* Efficient isolation and gene expression profiling of small numbers of neural crest stem cells and developing Schwann cells. *J. Neurosci.* **24,** 2357–65 (2004).

106. Engreitz, J. M., Daigle, B. J., Marshall, J. J. & Altman, R. B. Independent component analysis: mining microarray data for fundamental human gene expression modules. *J. Biomed. Inform.* **43,** 932–44 (2010).

107. Zilliox, M. J. & Irizarry, R. a. A gene expression bar code for microarray data. *Nat. Methods* **4,** 911–3 (2007).

108. Yang, X., Bentink, S., Scheid, S. & Spang, R. Similarities of ordered gene lists. *J. Bioinform. Comput. Biol.* **4,** 693–708 (2006).

109. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S. & Ebert, B. L. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. (2005).

110. Oldham, M. C. *et al.* Functional organization of the transcriptome in human brain. *Nat. Neurosci.* **11,** 1271–82 (2008).

111. Repsilber, D. *et al.* Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics* **11,** 27 (2010).

112. Gaujoux, R. & Seoighe, C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study. *Infect. Genet. Evol.* **12,** 913–921 (2012).

113. Gaujoux, R. & Seoighe, C. CellMix: A comprehensive toolbox for gene expression deconvolution. *Bioinformatics* **29,** 2211–2212 (2013).

114. Vahedi, S. *et al.* Parkinson's disease candidate gene prioritization based on expression profile of midbrain dopaminergic neurons. *J. Biomed. Sci.* **17,** 66 (2010).