

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Unbiased Next Generation Sequencing Assays Illuminate Idiopathic Meningitis and Encephalitis

**Permalink**

<https://escholarship.org/uc/item/9fg6k3b1>

**Author**

O'Donovan, Brian Daniel

**Publication Date**

2018

Peer reviewed|Thesis/dissertation

Unbiased Next Generation Sequencing Assays  
Illuminate Idiopathic Meningitis and Encephalitis

by

Brian D. O'Donovan

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

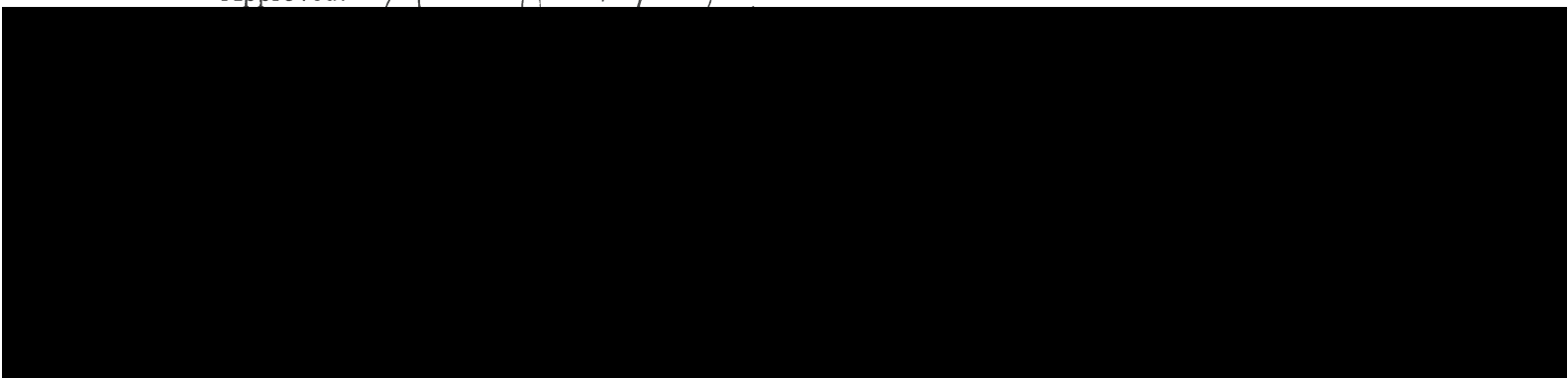
in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:



Copyright 2018

by

Brian D. O'Donovan

## **ACKNOWLEDGEMENTS and DEDICATION**

There are several people in my life whom, without their support and encouragement, I would not have had the courage, means, or stamina to complete this endeavor. Foremost is my advisor and mentor, Dr. Joe DeRisi. Joe has seemingly inexhaustible reserves of energy, excitement, wonder, and knowledge. His open door, open access, unencumbered approach to science is as inspiring as it is effective. For all the self-doubt, second guessing, and anxiety that comes with pursuing a PhD, I never had any uncertainty about my decision to join the DeRisi Lab.

There are two people that I consider my unofficial and uncredited co-mentors - Eric Chow and Michael Wilson. Eric embodies all of the positive and noble aspects of academia I imagined when I first made the decision to apply to graduate school. His generosity with his time and expertise is unparalleled and UCSF is a better place for him being there. Much of this work (and many other graduate theses, I'm sure) would not have been possible without his personal and professional stewardship. Michael's combination of intellect, charm, and humility would be irritating if it weren't so genuine. He possesses a temperament and work ethic that make all of his remarkable accomplishments seem effortless. I don't think I have met anyone as perfectly suited for their position in life; he is the consummate physician, scientist, friend, and an amazing father to his children. His work in idiopathic encephalitis and meningitis is the genesis of a majority of my graduate work and I consider him a personal and professional role model. We will take that fishing trip, Michael.

I'd like to recognize my thesis committee, Drs. Katie Pollard, Jim Wells, and Sam Pleasure. I didn't know it at the time, but I was assembling a veritable dream team of



players if the purpose of the game was casual and enlightening conversations about my work, my future plans, and what I needed to be a successful scientist. Committee meetings provided an annual boost of confidence and excitement about my work that, after being so close to it on a daily basis could start to feel pedestrian or somehow lacking. I can't take credit for choosing such an ideal committee but I am enormously proud that it will be their signatures on this document.

I was accepted into a remarkable cohort of fellow graduate students and postdocs, many I now call close friends. I'd like to acknowledge John Hawkins, Brian Sharon, Kieran Mace, Garrett Gaskins, Clint Cario, Veronica Pessineo, and Christina Homer for simply being who they are. Within the DeRisi Lab in particular, I'd like to recognize Greg Fedewa, Christine Sheridan, Vida Ahyong, Hanna Retallack, Emily Crawford and Katrina Kalantar. Thanks for putting up with me and sharing these last 5 years of long hours, botched experiments, microwaved fish, and questionable leftovers.

Finally, my family. Any modicum of success I have or will achieve in life can be directly attributed to the unwavering and unconditional support I get from my mother, Noreen Coakley. She is one of the most formidable and impressive people I've ever met. Her drive, tenacity, and belief in the liberating power of education are what have afforded me this life of financial and intellectual freedom. Not everyone can decide to upend their lives and spend their early thirties playing with computers and viral genomes. Thanks, mom. My father, Bert, died suddenly just as my graduate school adventure was winding down. While I always got the impression that my decision to commit to another 6 years of school was discordant with his pragmatic, tradesman sensibilities, I never felt anything but support, encouragement, and pride coming from

him. I always looked forward to summarizing my research in a final exit talk that struck the perfect tone and was geared for any audience. I realize now that in my mind's eye, he was that audience. My brother, Dominic, is perhaps the most potent antidote to an inflated ego and the trappings of self-importance that doesn't require illicit chemical synthesis. His influence over the last 33 years is the only reason this "nerdbot" can pass the Turing test in most social situations. He will probably never realize or admit the extent or depth of his influence or the deftness with which he wields it.

Above all, I'd like to dedicate this work to my wife and partner, Danielle. It's been a crazy 11 years. You lapped me in getting the doctorate, some might claim you are funnier, and you put up with me during my most neurotic and anxious episodes with grace and levity. Here's to many more adventures. Now that this is all over I can finally focus on writing you that chart topper.

## CONTRIBUTIONS

Chapter 3 of this dissertation contains reprinted or adapted material and figures from the previously published manuscript:

*Wilson MR, O'Donovan BD, Gelfand JM, Sample HA, Chow FC, Betjemann JP, Shah MP, Richie MB, Gorman MP, Hajj-Ali RA, Calabrese LH, Zorn KC, Chow ED, Greenlee JE, Blum JH, Green G, Khan LM, Banerji D, Langelier C, Bryson-Cahn C, Harrington W, Lingappa JR, Shanbhag NM, Green AJ, Brew BJ, Soldatos A, Strnad L, Doernberg SB, Jay CA, Douglas V, Josephson SA, DeRisi JL. "Chronic Meningitis Investigated via Metagenomic Next-Generation Sequencing." JAMA Neurology (2018)*

Michael Wilson and Joe DeRisi conceived of and funded the study and edited the manuscript. Michael Wilson generated a majority of the sequencing libraries is responsible for the text describing the clinical presentation and workup for each patient.

Chapter 4 contains reprinted or adapted material and figures from the previously published manuscript:

*Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, DeRisi JL. "Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications." Genome Biology (2016)*

Wei Gu and Emily Crawford performed experiments. Emily Crawford was responsible for a majority of the manuscript introduction and methods.

# **Unbiased Next Generation Sequencing Assays Illuminate Idiopathic Meningitis and Encephalitis**

Brian D. O'Donovan

## **ABSTRACT**

Encephalitis and associated conditions (meningitis, myelitis, etc.) are a collection of diseases characterized by acute or chronic inflammation of the brain and central nervous system (CNS). Diagnosing and treating encephalitis can be challenging given the litany of potential infectious, ischemic, metabolic, autoimmune, neoplastic, paraneoplastic, parameningeal and toxic causes. Diagnostic approaches in infectious and autoimmune etiologies – which account for an overwhelming majority of cases – are generally candidate-based panels of nucleic acid (PCR), serological, or cell-based assays with limited throughput and often suffer from low sensitivity and specificity. Frequently, patients receive no definite diagnosis, are treated empirically at great cost to the individual and greater health system, or receive treatments contraindicated to their specific condition. Novel, unbiased, and high-throughput diagnostic methods are needed to address these challenges.

Here I present my efforts to develop rapid, cost effective, unbiased, and high-throughput methods leveraging next generation sequencing (NGS) technologies to illuminate the underlying causes of infectious and autoimmune-mediated CNS

inflammation. I demonstrate the design and implementation of a phage display and immunoprecipitation (PhIP-Seq) assay to map the antigenic determinants of autoantigens in two paraneoplastic neurological disorders, the anti-Hu and anti-Yo syndromes. The assay, leveraging a rationally designed library of phage clones expressing >700,000 peptides encompassing the entire human proteome, also serves as a discovery platform to identify novel disease-associated antigens and antibody binding signatures that may eventually serve as valuable diagnostic and prognostic biomarkers in cancer detection and immune repertoire profiling.

To address challenges in identifying infectious etiologies, I present my efforts in applying metagenomic and metatranscriptomic NGS (mNGS) to patient cerebrospinal fluid (CSF) samples. The low nucleic acid content of CSF and the sterile, privileged nature of the CNS present unique opportunities and challenges for metagenomic assays. The complex, noisy, and polymicrobial datasets generated by mNGS require careful analysis to determine which, if any, of the identified microbes represent a true pathogen versus environmental contamination. Failure to make this distinction can result in spurious disease associations with organisms later determined to be laboratory contaminants. I demonstrate the necessity and efficacy of a straightforward statistical framework coupled with careful study design for identifying microbial pathogens in a series of challenging cases of subacute or chronic infectious meningitis, as well as for analyzing publicly-available data from recent mNGS diagnostic and brain microbiota studies.

# Contents

<b>Chapter 1 Background and Introduction.....</b>	<b>1</b>
1.1 Encephalitis and Meningitis – Prevalence and Diagnostic Challenges.....	1
1.2 Clinical Potential of Next Generation Sequencing (NGS) Technologies.....	2
1.3 Summary of Work and Findings.....	3
1.3.1 Autoimmune Encephalitis and PhIP-Seq.....	3
1.3.2 Infectious Encephalitis and Metagenomics.....	4
1.3.3 DASH.....	5
<b>Chapter 2 Exploration of Anti-Yo and Anti-Hu paraneoplastic neurological disorders by PhIP-Seq reveals a highly restricted pattern of antibody epitopes ....</b>	<b>6</b>
2.1 ABSTRACT.....	6
2.2 INTRODUCTION.....	7
2.2.1 Paraneoplastic Neurological Disorders.....	7
2.2.2 Anti-Yo PND.....	8
2.2.3 Anti-Hu PND.....	9
2.2.4 PhIP-Seq.....	9
2.3 RESULTS.....	11
2.3.1 Design, Production, and Characterization of PhIP-Seq Library.....	11
2.3.2 Fidelity of Library.....	12
2.3.3 Development of Statistical Approach to Analyze Peptide Enrichment.....	13
2.3.4 Phage Library Performance Validation: Commercial Antibodies.....	15
2.3.5 PND Cohort Results.....	16
2.3.6 Identification of CDR2/CDR2L antigens in anti-Yo samples.....	17

2.3.7	Identification of nELAVL antigens in anti-Hu samples .....	18
2.3.8	Consensus motif contains subsequences with predicted and experimentally demonstrated T-Cell Antigenicity .....	21
	PhIP-Seq Identifies Additional Known and Potentially Novel Autoantigens .....	23
2.3.9	Comparison to Healthy Sera Samples Identifies Potential Disease- Associated Binding Signatures .....	23
2.3.10	Zic and Sox Family Enrichment .....	24
2.3.11	Novel PND/Cancer Associated Signatures in Anti-Yo Patients .....	24
2.4	DISCUSSION .....	25
2.5	MATERIALS AND METHODS .....	30
2.5.1	Computational Design of Library .....	30
2.5.2	Cloning into T7 Select vector .....	31
2.5.3	Cloning and Packaging .....	32
2.5.4	In vivo amplification .....	33
2.5.5	Concentrating and Storing Phage Library .....	33
2.5.6	Immunoprecipitation .....	33
2.5.7	NGS Library Prep .....	34
2.5.8	Bioinformatic Analysis .....	36
2.5.9	Design of Hu Motif Mutational Scanning Library .....	37
2.5.10	ELAVL4 amplicon sequencing from bulk thymus and human TECs .....	38
2.5.11	Motif Discovery .....	40
2.5.12	Clinical PND Antibody Confirmation .....	41
2.6	FIGURES .....	42

2.7	TABLES .....	53
<b>Chapter 3 Metagenomic next-generation sequencing for chronic meningitis: a</b>		
<b>case series .....</b>		
		<b>55</b>
3.1	ABSTRACT .....	55
3.2	INTRODUCTION.....	56
3.3	RESULTS.....	57
3.3.1	Case Descriptions.....	57
3.3.2	Background Signature of Reagent and Environmental Contaminants .....	64
3.4	DISCUSSION AND CONCLUSIONS.....	65
3.5	MATERIALS and METHODS.....	68
3.5.1	mNGS Protocol.....	68
3.5.2	Bioinformatics and Statistical Analysis .....	69
3.6	FIGURES .....	71
3.7	TABLES .....	77
<b>Chapter 4 Depletion of Abundant Sequences by Hybridization (DASH): using Cas9</b>		
<b>to remove unwanted high-abundance species in sequencing libraries and</b>		
<b>molecular counting applications .....</b>		
		<b>78</b>
4.1	ABSTRACT .....	78
4.2	INTRODUCTION.....	79
4.3	RESULTS.....	82
4.3.1	Reduction of unwanted abundant sequences in HeLa samples.....	83
4.3.2	Enrichment of non-targeted sequences in HeLa samples .....	84
4.3.3	Reduction of unwanted abundant sequences in CSF samples .....	85



4.3.4	Reduction of wild-type background for detection of the KRAS G12D (c.35G>A) mutation in human cancer samples .....	86
4.4	DISCUSSION.....	88
4.4.1	Input requirements .....	90
4.4.2	Performance .....	90
4.4.3	Programmability.....	92
4.4.4	Cost.....	92
4.5	CONCLUSIONS.....	94
4.6	MATERIALS AND METHODS .....	95
4.6.1	Generation of cDNA from HeLa cell line and clinical samples.....	95
4.6.2	In vitro preparation of the CRISPR/Cas9 complex .....	95
4.6.3	CRISPR/Cas9 treatment.....	97
4.6.4	High-throughput sequencing and analysis of sequencing data .....	98
4.6.5	ddPCR of KRAS mutant DNA.....	99
4.6.6	Ethics .....	101
4.6.7	Availability of data and materials .....	101
4.7	FIGURES .....	102
4.8	TABLES .....	109
	<b>REFERENCES.....</b>	<b>110</b>

## List of Figures

Figure 2.1 - Design and Characterization of Library. ....	42
Figure 2.2 - Validation of library by commercial antibody IP's. ....	43
Figure 2.3 - Anti-Yo representative results. ....	44
Figure 2.4 - Anti-Yo epitope mapping.....	46
Figure 2.5 - Anti-Hu epitope mapping. ....	48
Figure 2.6 - Patient 01 samples dominated by ELAV motif.....	49
Figure 2.7 - Anti-CRMP5 epitope mapping.....	50
Figure 2.8 - Scaling and normalizing peptide fold-change values.....	51
Figure 2.9 - PhIP-Seq Identifies short exon in ELAVL4 excluded in thymus.....	52
Figure 3.1 - Diagram of a rapid computational pipeline. ....	71
Figure 3.2 - Ranked Results of Statistical Scoring. ....	72
Figure 3.3 - Selected Neuroimaging. ....	73
Figure 3.4 - Background signature of reagent and environmental contaminants. ....	74
Figure 3.5 - RNA doping experiment. ....	76
Figure 4.1 - DASH conceptual figure .....	102
Figure 4.2 - DASH targeting abundant mitochondrial rRNA in Hela extractions.....	103
Figure 4.3 - DASH'ed clinical samples. ....	104
Figure 4.4 - DASH in cancer.....	105
Figure 4.5 - Concentration-dependent depletion. ....	106
Figure 4.6 - Mitochondrial rRNA target sites in mtRNR2L12 .....	107
Figure 4.7 - Patient-specific scatterplots.....	108

## List of Tables

Table 2.1 - Significant antigen-disease associations. ....	53
Table 2.2 - Preferred <i>E. coli</i> codons used in library design. ....	54
Table 3.1 - Clinical characteristics of study participants.....	77
Table 3.2 - Metagenomic sequencing summary .....	77
Table 4.1 - Summary of depletion/enrichment results in DASH-treated CSF .....	109
Table 4.2 - sgRNA targeted sequences .....	109

# Chapter 1

## Background and Introduction

### 1.1 Encephalitis and Meningitis – Prevalence and Diagnostic Challenges

Encephalitis, meningitis, and many closely associated conditions (myelitis, CNS vasculitis, etc.) are a set of diseases characterized by acute or chronic inflammation of the brain and central nervous system (CNS). In the US, encephalitis accounts for over 20,000 hospitalizations annually with a mortality rate of 6% and estimated in-patient cost of \$2 billion.<sup>1,2</sup> A major contributor to this disease burden is the fact that CNS inflammation presents a unique diagnostic challenge given the litany of potential infectious, ischemic, metabolic, autoimmune, neoplastic, paraneoplastic, parameningeal and toxic causes. Diagnostic approaches in infectious and autoimmune etiologies – which account for an overwhelming majority of cases – are generally low-throughput, candidate-based panels of nucleic acid (PCR), serological, or cell-based assays (CBAs) and often suffer from low sensitivity and specificity. As many as 40-70% of patients receive no definite diagnosis,<sup>2</sup> are treated empirically at great cost to the individual and greater health system, and can receive treatments contraindicated to their specific condition (immune suppression during viral infections, for example) as physicians act in

good faith but with limited information. Novel, unbiased, and high-throughput diagnostic methods are needed to address these challenges and better understand the underlying mechanisms that lead to CNS inflammation.

## **1.2 Clinical Potential of Next Generation Sequencing (NGS) Technologies**

Next Generation Sequencing (NGS) describes a suite of platforms and technologies defined by the ability to sequence millions of small DNA fragments in parallel.<sup>3</sup> The impact of NGS on molecular biology research and our understanding of the natural world cannot be understated. NGS has ushered in a revolution in biomedical research and genomic data is estimated to become the largest source of digital information in world by 2025.<sup>4</sup> Despite its relatively ubiquitous station in the basic research laboratory, adoption and implementation of NGS technologies in the clinical context has been slower. This has been in part due to the cost and availability of sequencing, accompany data storage and compute requirements, and the appropriate regulatory due diligence necessary in adopting new methodologies in the higher-stakes medical field.

Major declines in sequencing costs, distributed (“cloud-based”) compute and data storage solutions, and increased availability and familiarity with the technology in medical community has enabled NGS to make substantial inroads in recent years. The growing consensus is that NGS is poised to precipitate a similar revolution in medicine, having already made significant contributions in prenatal testing,<sup>5</sup> cancer screening and genotyping,<sup>6</sup> and infectious disease diagnosis and monitoring.<sup>7</sup> Integration of whole exome or single nucleotide polymorphism (SNP) panel data with electronic medical

records (EMRs) will enable novel disease-genotype associations, drug response prediction and stratification, and reveal novel, polygenic disease phenotypes that have thus far eluded clinical understanding. Further, NGS will continue to precipitate discoveries in basic research that will expand our understanding of the molecular basis for disease and undoubtedly have translational implications.

In this dissertation, I present my work and efforts in applying NGS technologies to developing rapid, cost-effective, high-throughput assays for the diagnosis and understanding of autoimmune and infectious causes of encephalitis. While each chapter is meant to be self-contained or is based on previously published material, a brief summary of the history, motivation and context of each chapter/project is in order.

### **1.3 Summary of Work and Findings**

#### **1.3.1 Autoimmune Encephalitis and PhIP-Seq**

Autoimmune encephalitis describes a disease process wherein the patient's own immune system fails to maintain the ability to distinguish "self" from "non-self" and the arsenal of cellular and molecular sentinels that normally protect the body from infection or malignant neoplasm are brought to bear against the patient's CNS cells and tissues. The mechanisms leading to autoimmunity vary are not fully understood, and the array of potential endogenous target proteins, known as autoantigens, is as heterogeneous as the human proteome. In chapter 2 of this work, I describe the design and implementation of a programmable phage display library and automated, robotics-based immunoprecipitation and sequencing protocol (PhIP-Seq) that allows the simultaneous screening of dozens of patient samples for antibodies that bind to all possible

endogenous peptide antigens. The data provides a high-resolution snapshot of antibody binding signatures and can map antigenic determinants with single amino acid resolution. I apply PhIP-Seq to two common paraneoplastic conditions, the anti-Hu and anti-Yo syndromes, though the technique is applicable to any autoimmune disorder. I characterize the binding signatures to the canonical Yo and Hu antigens while identifying several known and novel disease-associated autoantigens. The data provides insight into the oncoimmune processes underpinning the rise of autoimmunity and identifies several potentially novel diagnostic and prognostic biomarkers in paraneoplastic encephalitis and cancer.

### **1.3.2 Infectious Encephalitis and Metagenomics**

The CNS has unique anatomic and immunologic characteristics that make it particularly vulnerable to infection from foreign microbes.<sup>8</sup> Despite ostensible protection from the “blood-brain barrier,” hundreds of environmental and commensal bacteria, viruses, fungi, protozoa, helminths and eukaryotic parasites are known to infect all levels and anatomic structures of the CNS - proliferating with impunity in the immunologically “privileged” environment.<sup>9</sup> Roughly half of all hospitalizations related to CNS inflammation are considered to be infectious, though limitations of existing diagnostic assays mean that often no etiologic pathogen is ever identified.<sup>2</sup>

Metagenomics is the study of all genetic material and associated taxa in a biological sample. Using metagenomic NGS (mNGS) to identify pathogens in infectious disease is perhaps the most salient clinical application of NGS technology. In chapter 3 of this dissertation, I outline my efforts to apply mNGS to cerebrospinal fluid samples

(CSF) and develop statistical approaches and study design standards to identify the causative pathogen in a series of acute and chronic cases of infectious meningitis. I demonstrate and emphasize the need for robust statistical and methodological approaches in analyzing the inherently complex and polymicrobial datasets generated by mNGS to avoid false positives and potential misdiagnosis. I apply my approach to publicly available data from recent publications implicating common laboratory reagent contaminants in CNS infection or purporting the existence of a brain or CNS microbiome.<sup>10</sup>

### **1.3.3 DASH**

In analyzing hundreds of clinical mNGS datasets in the DeRisi lab, it became apparent that only a minute fraction (often <0.1%) of the sequencing data “reads” were derived from causative pathogen. A huge majority of reads were attributed to a small number of ubiquitously expressed human mitochondrial genes that were filtered out and otherwise ignored in our computational pipeline. These reads consume valuable storage and compute space and increase the requisite depth of sequencing required to detect the low abundant species in a sample. In chapter 4, I describe a simple and programmable biochemical method devised to remove unwanted sequences in oligonucleotide libraries just prior to loading on the sequencing platform. The method, dubbed Depletion of Abundant Sequences by Hybridization (DASH), leverages the programmable and specific exonuclease activity of the Cas9 protein and proved to be an easy, cost efficient, and highly effective solution to a problem that plagues many NGS applications.



## Chapter 2

# Exploration of Anti-Yo and Anti-Hu paraneoplastic neurological disorders by PhIP-Seq reveals a highly restricted pattern of antibody epitopes

### 2.1 ABSTRACT

Paraneoplastic neurological disorders (PNDs) are immune-mediated diseases of the nervous system understood to manifest as part of a misdirected anti-tumor immune response. Identifying PND-associated autoantibodies and their cognate antigens can assist with proper diagnosis and treatment while also enhancing our understanding of tumor-associated immune processes, triggers for autoimmune disease, and the functional significance of onconeural proteins. Here, we employed an enhanced version of phage display immunoprecipitation and sequencing (PhIP-Seq) leveraging a library of over 731,000 unique phage clones tiling across the entire human proteome to detect autoantibodies and create high-resolution epitope profiles in serum and CSF samples from patients suffering from two common PNDs, the anti-Yo (n = 36 patients) and anti-Hu syndromes (n = 44 patients). All patient samples positive for anti-Yo antibody by a validated clinical assay yielded polyspecific enrichment of phage

presenting peptides from the canonical anti-Yo (CDR2 and CDR2L) antigens, while 38% of anti-Hu patients (17/44) had a serum and/or CSF sample that significantly enriched peptides deriving from the ELAVL family of proteins, the anti-Hu autoantigenic target. The anti-Hu antibodies showed a remarkably convergent antigenic signature across 15/17 patients corresponding to residues surrounding and including the degenerate motif, RLDxLL, shared by ELAVL2, 3 and 4. Lastly, PhIP-Seq identified several known and novel autoantigens in these same patient samples, representing potential biomarkers that could aid in the diagnosis and prognosis of PND and cancer.

## **2.2 INTRODUCTION**

### **2.2.1 Paraneoplastic Neurological Disorders**

Paraneoplastic neurological disorders (PNDs) affecting the central nervous system (CNS) are a collection of diseases characterized by an autoimmune response to proteins nominally restricted to the CNS and triggered by ectopic expression of the antigen by a systemic neoplasm.<sup>11</sup> Neurologic symptoms of PNDs vary and can reflect damage to essentially any level of the neuroaxis. This symptom variability is, in part, determined by the neuroanatomic and/or cell type specificity of each PND. Existing PND pathogenesis models posit that an otherwise effective anti-tumor immune response to tumor antigens precipitates a breakdown of immune tolerance and subsequent infiltration of autoreactive lymphocytes and/or pathogenic autoantibodies into the CNS.<sup>12</sup> The roles and culpability of cell-mediated and humoral immune responses in disease progression are disputed, as onconeural antibodies are often specific for intracellular proteins and are frequently detected in the sera of cancer patients in the absence of

PND symptoms.<sup>13</sup> In a typical clinical presentation, the neurological deficits and symptoms manifest before any underlying cancer is detected. As such, patients presenting with PND symptoms and tumor-associated antibodies are presumed to have cancer until otherwise ruled out and long-term tumor monitoring is recommended. Patients developing PND symptoms have lower rates of metastasis and higher rates of cancer survival, likely a consequence of the otherwise appropriate oncoimmune response. Additionally, a majority of PND patients test positive for multiple tumor/PND related autoantigens.<sup>14</sup> Antibody specificities are thought to be reflective of a tumor's onconeural antigen expression, and clusters of neural autoantibodies in a patient's serum or CSF can aid in specific cancer diagnosis.<sup>15</sup>

### **2.2.2 Anti-Yo PND**

Anti-Yo paraneoplastic cerebellar degeneration (PCD) accounts for roughly 50% of subacute PCD presenting in middle age.<sup>16</sup> The disease is marked by severe cerebellar ataxia, dysarthria, and Purkinje cell death and almost exclusively affects women, as it is primarily associated with gynecological and breast malignancies.<sup>17</sup> Anti-Yo antibodies, also known as anti-Purkinje cell cytoplasmic antibody 1 (PCA-1), target the two intracellular proteins CDR2 (cerebellar degeneration related protein 2) and/or CDR2L (CDR2-like). Little is known about the function of these proteins other than that they have a role in transcription and that their expression is primarily limited to Purkinje neurons in the cerebellum, the brainstem, and spermatagonia.<sup>18</sup> Whether the antibodies alone have pathogenic potential remains controversial; however, it is generally accepted that specific CD8+ T cells lead to Purkinje cell death.<sup>19</sup> Although

disease prognosis is poor, early detection, immunosuppression, and successful cancer treatment can limit Purkinje neuron loss and improve patient outcomes.<sup>20</sup> To date, the specific antigenic determinants on the CDR2 and CDR2L proteins have not been mapped.

### **2.2.3 Anti-Hu PND**

Anti-Hu-related PND is most commonly associated with small cell lung cancer (SCLC) and can present with differing neurologic syndromes ranging from dysautonomia, encephalopathy (limbic, brainstem, multifocal), epilepsy, cerebellar degeneration, myelopathy, and polyneuropathy.<sup>21</sup> Anti-Hu antibodies target members of the ELAVL (embryonic lethal abnormal vision like) family of intracellular RNA-binding proteins. There are 4 members of this family, ELAVL1, ELAVL2, ELAVL3, and ELAVL4 (also known as HuA/R, HuB, HuC, and HuD, respectively). ELAVL1 is the most divergent homolog and is ubiquitously expressed in all tissues, while ELAVL2, 3, and 4 have restricted expression to the nervous system and are often referred to as nELAVL proteins (neuronal ELAVL).<sup>22</sup>

### **2.2.4 PhIP-Seq**

The majority of currently available clinical autoimmune panels include antigen-specific ELISA-based methods, cell-based assays (CBAs), co-localization immunohistochemistry (IHC), and traditional western blots.<sup>23</sup> These candidate-based or single antigen tests may not capture the true complexity of the presumably more heterogeneous oncoantigenic repertoire. Unbiased and higher throughput methods exist

in the research context and include immunoprecipitation-mass spectrometry (IP-MS), expression library screening, biochemical purification, peptide arrays, protein arrays, and phage display, each with its own advantages and limitations. Among these, variations on bacteriophage display technologies have been successfully utilized for decades in antigen discovery,<sup>24</sup> epitope mapping,<sup>25</sup> and antibody engineering.<sup>26</sup> Traditionally, tissue-specific cDNA libraries or degenerate or random synthetic oligonucleotides are cloned and expressed on the surface of phage capsids which are panned against immobilized antibody or other targets of interest. More recently, rationally designed libraries encompassing the entire human proteome have been implemented.<sup>27</sup> With next generation sequencing (NGS) as a readout, researchers can quantify the enrichment of millions of individual phage clones simultaneously and identify sequences that bind to the target or antibody of interest. This technology, known as Phage Immunoprecipitation and Sequencing (PhIP-Seq), enables high-throughput screening of patient samples and has successfully identified autoantigens and mapped epitopes in inclusion body myositis, rheumatoid arthritis, as well as two other common PNDs, the anti-Ma and anti-Ri syndromes.<sup>28</sup> While broad spectrum autoantibody screening has previously been considered to be cost prohibitive,<sup>29</sup> the continuing decline of large scale oligonucleotide synthesis and NGS costs may make whole proteome screening in by PhIP-Seq increasingly accessible in both research and clinical contexts.

Inspired by the potential of this technology to build upon our existing studies into the underlying causes of meningitis and encephalitis, we expanded on the published PhIP-Seq technique by designing a larger, more comprehensive library of endogenous

human peptides. Our human “peptidome” consists of 731,724 unique phage clones and includes all known and predicted human protein splice variants. Using this expanded library, we conducted nearly 800 immunoprecipitations over 130 serum and CSF samples from patients suffering from anti-Hu and anti-Yo PNDs, revealing the most comprehensive and highest resolution epitope mapping of the autoantigens targeted in these PNDs to date. For the unusually convergent Anti-Hu epitope, additional deep-mutational scanning phage libraries were employed to map binding determinants at the single amino acid level. The results presented here further demonstrate the utility and high-resolution capability of programmable phage display for both basic science and clinical applications.

## **2.3 RESULTS**

### **2.3.1 Design, Production, and Characterization of PhIP-Seq Library**

The Human PhIP-Seq Library v2 contains all annotated human protein sequences from the NCBI protein database as of November 2015, including all published and computationally predicted splice variants and coding regions (Figure 1). Full-length sequences were clustered on 99% sequence identity (CD-HIT v4.6)<sup>30</sup> to remove duplicate and partial entries, resulting in a set of 50,276 proteins. Each protein was computationally divided into 49 AA peptides using a 24 AA sliding window approach such that sequential peptides overlapped by 25-residues (see Methods). This resulted in a set of 1,256,684 peptide sequences encompassing the entire human proteome. To remove redundant sequences derived from identical regions of the various isoforms and homologs, we further clustered and collapsed the 49 AA peptides

on 95% identity such that peptides with 2 or fewer AA differences were combined by choosing one representative sequence.

Amino acid sequences were converted to nucleotides using preferred *E. coli* codons (see Methods).<sup>31</sup> Restriction sites (EcoRI, HindIII, BamHI, XhoI) were removed with synonymous nucleotide changes to facilitate cloning. To enable amplification from synthesized oligo pools, 21 nucleotide (nt) universal priming sites were added to the 5' and 3' ends of each sequence. These priming sites also encoded STREP and FLAG tag sequences (after addition of a terminal codon by PCR). Thus, the final library consisted of 731,724 unique 189mer oligonucleotide sequences encoding 49 AA human peptides flanked by 5'-STREP and 3'-FLAG sequences. The library was synthesized by Agilent Technologies (Santa Clara, CA) in three separate pools. The complete sequence of each oligo is available at <https://github.com/derisilab-ucsf/PhIP-PND-2018>.

### **2.3.2 Fidelity of Library**

After PCR amplification and size selection (Blue Pippin 3% Agarose cassette, Sage Science, Beverly, MA) but prior to cloning and packaging into phage, the synthesized oligos were evaluated by NGS with 70 million paired-end 125 nt reads (Figure 1B). 98.5% of all the synthesized sequences were full length, and 84% were error free. For those with errors, a majority (10% of all sequences) were deletions of 3 or fewer bases, while single base substitutions accounted for 4%, and the remaining 2% of oligos contained larger deletions (> 4 bases) or truncated/chimeric sequences. The commercially synthesized library yielded 722,436 unique peptide sequences with an estimated Chao1 diversity of 723,421, indicating that 99% of the library was present.

The distribution and coverage of the library was uniform, with 99.9% of the sequences within ten-fold of the expected fraction (Figure 1C).

Size-selected and restriction-digested oligos were cloned and packaged in 25 separate 10  $\mu$ l reactions according to the manufacturer's specifications (EMD, Burlington, MA). The efficiency of *in vitro* packaging was quantified by plaque assay. The diluted and quenched packaging reaction (3 ml total) contained  $10^9$  plaque forming units (pfu)/ml, roughly 1000x coverage of the entire library. The library of phage clones was sequenced (70 million paired-end 125 nt reads) to determine the fidelity of the packaged library. 77% of phage sequenced yielded error-free, full-sized inserts (Figure 1). The majority of errors (21% of all sequences) were deletions or stop codons resulting in shorter expressed peptides. Within the 77% of correct sequences, there were 657,948 unique clones with an estimated Chao1 diversity of 658,476 – indicating that 89% of this library was packaged and cloned without any errors. Allowing for synonymous mutations and single amino acid substitutions, 92% of the library was packaged successfully into phage.

### **2.3.3 Development of Statistical Approach to Analyze Peptide Enrichment**

Several statistical approaches for analyzing PhIP-Seq data have been published, primarily based on Generalized Poisson (GP) or Negative Binomial (NB) count models fit to the distribution of unselected or input library phage populations.<sup>32,33</sup> More recently, an algorithm leveraging z-scores of phage counts relative to AG bead only “mock” IPs was developed to account for non-specific phage binding to IP reagents and substrates.<sup>34</sup> These approaches were developed to analyze datasets generated from



single rounds of immunoprecipitation and were optimized for sensitivity to low abundance phage.

In order to identify phage significantly enriched after multiple consecutive rounds of selection with patient antibodies, a statistical framework was developed (see Methods) based on phage fold-change statistics relative to a collection of 40 independent mock IPs using protein-AG beads alone. Briefly, peptide counts from each control and sample IP were normalized to reads per 100,000 (rpK). Normalized peptide counts from each experimental IP were multiplied by a sample-specific scaling factor derived from the median rpK of the 100 most highly and consistently enriched AG-bead binding peptides (Figure 2.8, sequences available at: <https://github.com/derisilab-ucsf/PhIP-PND-2018>). Each peptide's fold-change enrichment was then calculated by dividing its scaled rpK value by its mean representation in the AG controls. Given the nature of the fold-change calculation, small fluctuations in low abundance phage in control IPs can lead to inflated fold-change values and false positives. To address this, fold-change values were further transformed using a method inspired by smoothed quantile normalization approaches in microarray data (see Methods).<sup>35</sup> P-values were assigned to the fold-changes by evaluating the survival function of a normal distribution fit to the  $\log_{10}(\text{fold-change})$  values in each experimental batch. Peptides reported as significant were those having an adjusted p-value ( $p_{\text{adj}}$ )  $\leq 0.05$  in at least two replicates after multiple test correction (Benjamini-Hochberg).<sup>36</sup>

### 2.3.4 Phage Library Performance Validation: Commercial Antibodies

Validation of the PhIP-Seq protocol was performed using two commercial antibodies with known specificities: anti-glia1 fibrillary acid protein (GFAP) (Agilent/DAKO, Carpinteria, CA) and anti-gephyrin (GPHN) (Abcam, Burlingame, CA) polyclonal antibodies. Three rounds of enrichment were used to further select for the particular epitope and reduce the proportion of non-specific binders. After each round of enrichment, phage populations were assessed by sequencing to an average depth of 2 million paired-end 125 nt reads. After 3 rounds of enrichment, a majority (50-70%) of the phage in the final population encoded peptides derived from the commercial antibody target (Figure 2.2A). Replicate IPs were conducted on separate days and exhibited high reproducibility ( $r > 0.8$ ) in phage counts between experiments.

When applied to these data, our statistical model correctly identified peptides from the target gene of interest as the most significantly enriched with minimal off-target or unrelated peptides/genes (Figure 2.2B). The anti-GFAP IPs enriched 18 unique peptides representing  $> 60\%$  of all phage in the final library (in both replicates) and scaled fold-changes greater than 300,000x ( $p_{\text{adj}} < 10^{-20}$ ) over the mock IPs. Similarly, the anti-GPHN commercial antibody enriched 4 unique C-terminal gephyrin peptides with 10,000-fold enrichment ( $p_{\text{adj}} < 10^{-25}$ ) over control IPs.

Alignment of the significantly enriched peptides from the anti-GFAP IP to the full-length protein revealed a distinct antigenic profile (Figure 2.2C), suggesting the commercial antibodies had highest affinity for a ~27 AA region surrounding the exon 5-6 junction. Notably, this region was identified by the immune Epitope Database (IEDB) tools Emini Surface Exposure calculation as having the highest B cell antigenicity

potential.<sup>37</sup> The three significantly enriched GPHN peptides were all overlapping sequences with consensus at the final 32 residues of the surface exposed C-terminus (Figure 2.2D). The anti-GPHN IPs also consistently enriched a single peptide from chromogranin A (CHGA) at levels greater than or commensurate with that of the GPHN peptides ( $p_{\text{adj}} < 10^{-30}$ ). Motif analysis (GLAM2)<sup>38</sup> and alignment of the CHGA and GPHN phage-peptides, revealed a discontinuous 6-residue sequence (YxE<sub>xx</sub>K) shared between CHGA and GPHN. Of the 256 peptides reported as significant in both replicates, 144 peptides (56%), contained the motif (Figure 2.2D). The probability of recovering these motif-containing peptides by chance given their representation in the original library is infinitesimal ( $< 10^{-300}$ , Fischer's exact test), indicating that a majority of the peptides in the final phage population were, in fact, true binders to the commercial antibody. These data also demonstrated that single AA resolution binding motifs could be recovered given sufficient representation in the original library.

### **2.3.5 PND Cohort Results**

In total, 798 individual IPs (including all three rounds of enrichment) were performed on 130 CSF and serum samples from patients with anti-Yo or anti-Hu antibodies identified by the Mayo Clinic Neuroimmunology Laboratory with CLIA and New York State approved methodology, as previously described (see Methods).<sup>39</sup> The PhIP-Seq protocol was performed on a Biomek FX (Beckman Coulter, Brea, CA) automated liquid handler across three experimental batches (see Methods). All samples were run blinded and in duplicate. A sample was reported as positive if peptides derived from the respective Yo (CDR2/CDR2L) or Hu (ELAVL2,3,4) antigens showed significant

fold-changes ( $p_{\text{adj}} \leq 0.05$ ) in both replicates. Files containing the peptide counts for each IP are available at <https://github.com/derisilab-ucsf/PhIP-PND-2018>, and raw data are available under NCBI BioProject Accession PRJNA506115.

### 2.3.6 Identification of CDR2/CDR2L antigens in anti-Yo samples

From 36 confirmed anti-Yo patients, a total of 53 samples were interrogated by PhIP-Seq, of which 51 (96%) were positive for anti-Yo peptides. When sample identities were unblinded, it was revealed that all 36 patients were represented by these positive samples. No significant differences were observed in sensitivity or binding profiles between CSF and serum, with 34/36 sera and 17/20 CSF testing positive ( $p_{\text{chi-square}} = 0.23$ ). From 50 de-identified healthy donor sera, only a single CDR2L peptide from a single donor was significantly enriched (37 rpK,  $p_{\text{adj}} = 0.0494$ ).

As shown in Figure 2.3A for three representative patients, phage specific for CDR2/CDR2L dominated the population by the final round of selection in a majority of samples. All patients showed enrichment for CDR2L while 30/36 (83%) were positive for both CDR2L and CDR2 peptides. No patients showed reactivity to CDR2 peptides alone. Across all patients, 29 unique CDR2L peptides were identified and accounted for 21% of all phage sequenced. Conversely, 6 unique peptides from CDR2 were identified across all patient samples and accounted for 0.25% of all phage sequenced (Figure 2.3B), suggesting that the overall magnitude and complexity of peptide enrichment for CDR2L was greater than for CDR2 peptides.

The antigenic profile across the full length CDR2L protein, for all 36 patients, is shown in Figure 2.4A. Convergent epitopes were shared by subgroups of patients,

particularly at residues 322-346, the most highly enriched epitope across all patients. In contrast to CDR2L, the antigenic profile of patient antibodies to CDR2 (Figure 2.4C) was primarily limited to 3 peptides spanning the 100 AA at the N-terminus. We further examined the relationship between homology and enrichment (Figure 2.4B).

Interestingly, the most highly enriched regions of CDR2L represented the most divergent regions between the two proteins, suggesting that the antibody responses to CDR2L and CDR2 are independent and patient antibodies are, in fact, preferentially enriching for CDR2L over CDR2.

### **2.3.7 Identification of nELAVL antigens in anti-Hu samples**

From 44 patients with laboratory confirmed anti-Hu PND, a total of 76 samples (32 paired serum/CSF, two CSF, and 10 serum) were subjected to PhIP-seq. A total of 19 samples from 13 patients resulted in significantly enriched peptides (in both replicates) derived from the nELAVL genes (ELAVL2, ELAVL3, and ELAVL4), the known targets of anti-Hu antibodies. Across these 19 samples, 20 unique nELAVL peptides were identified. While the majority (>80%) of enriched peptides were attributed to ELAVL4 (HuD). For the purpose of visualization, all nELAVL peptides from each sample were aligned to the most common isoform of ELAVL4 (accession NP\_001311142.1), shown in Figure 2.5A.

A majority (>90%) of the significantly enriched nELAVL peptides converged upon on a 17-residue sequence at AA positions 276-294 of ELAVL4, a sequence which is also common to variants of ELAVL2 and ELAVL3 (Figure 2.5B). The human peptidome PhIP-Seq library contained 18 peptides covering at least 10-residues of this short

sequence, reflecting the large number of published isoforms in the NCBI Protein database for these highly spliced and studied proteins. Due to this redundancy, rpK and fold-change calculations were attenuated by as much as two orders of magnitude as individual peptide counts were distributed across multiple sequences containing the same epitope. To account for this, the analysis was re-run treating all 18 peptides containing this short sequence as a single entity, the anti-Hu “signature.” This resulted in the detection of an additional 8 samples from 4 patients yielding significant enrichment of the signature sequence. Notably, a single patient (patient 26) enriched only a single peptide in both serum and CSF derived from the N-terminus of ELAVL2. The peptide is unique to ELAVL2 isoforms that had only been computationally predicted at the time of library design. These isoforms have since been confirmed experimentally and recognized as isoform c (accession NP\_001338384.1).<sup>40</sup> Importantly, none of 50 healthy, de-identified, patient sera samples yielded significant enrichment of nELAVL-derived peptides.

The convergent 17-residue signature sequence common to the majority of patient samples lies within a functionally significant hinge region between two RNA binding domains (RRM2 and RRM3) and may contain nuclear localization and export sequences responsible for control of subcellular localization and neuronal differentiation.<sup>41</sup> Though exon and isoform naming conventions vary in the literature, this region is used to distinguish unique splice variants of ELAVL2, ELAVL3, and ELAVL4 (sv1, sv2, sv3) based on their inclusion or exclusion of relevant exons, specifically those commonly referred to as exons 6 and 7a. The identified motif terminates at the exon 6/7a junction (Figure 2.5B). The majority of patient samples preferentially enriched

peptides that span the exon 6/7 junction, but do not include 7a, strongly suggesting that the binding determinants for the anti-Hu antibodies lie within this critical 17-residue anti-Hu signature region.

To further characterize the precise determinants of antibody binding, a new deep mutational scanning phage library was designed, encoding all possible single point mutants across the 17-residue signature motif (oligo design files available at <https://github.com/derisilab-ucsf/PhIP-PND-2018>). Samples from patients 01 and 38 were profiled by deep-mutational scanning PhIP-seq, as they represented the samples with the most and least complex peptide representation at this region, respectively. At each position, the fold-change enrichment of each amino acid substitution against an AG-only IP and relative to the abundance of the reference sequence is shown in Figure 2.5C, providing a landscape of the mutational tolerance at each residue. Examination of mutations that result in STOP codons (denoted by \* in figure) are particularly informative, revealing the minimal sequence length required for antibody binding. Patient 01 antibodies required a minimal epitope containing residues up to and including positions 245-255 (QAQRFRLDNLL) with a shorter subsequence (QRFRLDNLL) being least tolerant of mutation. Patient 38's antibodies appear to target a closely overlapping motif, RLDNLLN-AYG, with residues downstream not influencing peptide enrichment.

In light of this refined anti-Hu signature sequence, patient 01's original PhIP-Seq data was re-analyzed using GLAM2 (Gapped Local Alignment of Motifs) to identify enriched motifs. Of the 36 peptides identified as significant in both replicates, 34 (94%) share a short, 7-residue motif, identical to the RLDxxLL identified by the mutational scanning approach. These include 16 peptides that were not derived from nELAVL

genes, demonstrating that the dominant epitope in patient 01 is present in many unrelated protein sequences and explains the majority of peptides derived from non-nELAVL enriched genes. The biological significance of promiscuous binding to unrelated proteins harboring this motif is unknown.

### **2.3.8 Consensus motif contains subsequences with predicted and experimentally demonstrated T-Cell Antigenicity**

Examination of the refined anti-Hu signature motif and the full ELAVL4 sequence using the IEDB suite of sequence analysis tools (Figure 2.9) suggests that the motif has low predicted probability of being a linear B cell epitope (Bepipred 2.0). The motif does, however, contain several sequences predicted to have high affinity for major histocompatibility complex type 1 (MHC-I) alleles (A1, A2, and A3 supertypes) and high prediction scores for MHC processing and presentation, with two predicted proteasomal cleavage sites flanking the region at amino acid positions 246 and 264. Interestingly, peptides derived from this region have previously been identified as anti-Hu T cell antigens.<sup>42</sup> Indeed, peptides containing this exact motif have been shown to bind with high affinity to HLA-A1 (RLDNLLNMAY) and HLA-A2/B18 (NLLNMAYGV) and also to activate autologous CD8+ T cells obtained from Hu-positive patients.<sup>43</sup>

For paraneoplastic autoimmune diseases, the question of how immune tolerance is broken in the absence of new mutations remains to be addressed adequately. Selective exclusion of exons during expression in the thymus has been previously shown in mouse models as a possible route to break tolerance.<sup>44</sup> In the case of the anti-Hu ELAVL family, splice variants have been previously shown to be controlled by a concentration-dependent autoregulatory mechanism wherein ELAVL proteins bind to



AU-rich elements on nascent transcripts, protecting them from splicing and exclusion.<sup>45</sup> Therefore, it may be possible that protein levels in the thymus during the development of central tolerance might fail to reach the requisite levels, resulting in exclusion of certain exons which in turn could affect processing and presentation. To further investigate this speculation, publicly available RNAseq datasets derived from bulk and sorted murine thymus and thymic medullary epithelial cells (mTECs) were examined. While most datasets lacked sufficient nELAVL coverage necessary for analysis of splice variants, a single mTEC from a single cell RNAseq dataset contained >50X coverage across ELAVL4, revealing the exclusion of exon 7a (Figure 2.9).<sup>46</sup> Comparatively, sorted murine hippocampal and cerebellar neurons show essentially even coverage across all exons.<sup>47,48</sup>

To further investigate exon inclusion and exclusion of nELAVL splice variants during the development of central tolerance, amplicon libraries spanning exons 5, 6, 7a and 7 were generated from cDNA libraries derived from sorted human fetal and bulk adult thymic epithelial cells. An average of 1 million 125-nt paired-end reads (in triplicate) were obtained from each amplicon library. Exon 7a was represented in <0.5% of all reads in both fetal and adult TECs, supporting the notion that certain nELAVL exons, such as exon 7a, are largely excluded from thymic expression. Interestingly, exon 7a is adjacent to exon 6 containing the critical anti-Hu signature motif. Taken together, these data point to a potential mechanistic connection to central thymic tolerance whereby incomplete T cell tolerance to selective exons within nELAVL helps promote an autoantibody response against this region of the protein.

## **PhIP-Seq Identifies Additional Known and Potentially Novel Autoantigens**

Antigens, other than those classically thought to be anti-Hu or anti-Yo were also investigated. Anti-CRMP5 (DPYSL5/CV-2) antibodies are a well-established clinical biomarker for SCLC and malignant thymoma and can be found together with anti-Hu antibodies in patients with a PND.<sup>49</sup> Two of the anti-Hu patients (patients 51 and 52) had previously tested positive for anti-CRMP5 antibodies with a clinical assay (indirect immunofluorescence screening on mouse tissue substrate).<sup>50</sup> This was corroborated by PhIP-Seq, with both patients' CSF samples yielding significant enrichment of CRMP5 peptides. Patient 51 showed enrichment of 3 overlapping sequences at the C-terminus of the protein (AA residues 481-564) while patient 52's CSF enriched for two peptides spanning residues 73-217. Though not contiguous on the primary sequence, these regions of the protein are in contact on the surface of the published three-dimensional structure (Figure 2.6).

### **2.3.9 Comparison to Healthy Sera Samples Identifies Potential Disease-Associated Binding Signatures**

To further detect additional disease-associated antigens, the PND cohort was compared to the binding signatures derived from the sera of 50 de-identified healthy subjects, thus allowing detection of known or potentially novel antigens by virtue of their enrichment at levels significantly higher than the healthy. To simplify analysis and avoid assumptions about the congruence of specific peptide binding signatures across multiple patients, we collapsed the data to the gene-level and called a patient sample positive for a given gene if it yielded enrichment of any peptides for said gene in 2 or more samples/replicates. We compared the number of samples testing positive for each

gene in the PND groups vs the control and determined significance by Fisher's Exact test. The most significant gene-disease associations (FDR <0.05) are listed in Table 2.1. The full list of associations and healthy patient peptide counts are available at <https://github.com/derisilab-ucsf/PhIP-PND-2018>.

### **2.3.10 Zic and Sox Family Enrichment**

Among proteins that were significantly enriched, many have been previously identified as serological markers of cancer. Antibodies to ZIC (Zinc Fingers of the Cerebellum) and SOX (SRY-related HMG-box) families of transcription factors have a well-established clinical association with SCLC and PND.<sup>51,52</sup> Consistent with this fact, enrichment of peptides derived from the ZIC family (ZIC1, 2, 3, 4 and 5) was significant in 13/44 anti-Hu patients and 0/50 healthy controls ( $p_{\text{Fisher's exact}} = 0.000017$ ) while 19/44 patients and 2/50 healthy controls tested positive for SOX family peptides ( $p_{\text{Fisher's exact}} = 0.000004$ ). Though a majority of the peptides were derived from SOX1 and SOX2, the most common sequence, present in 13/19 patients with SOX reactivity, was a 34-residue sequence common to SOX1, 2, 3, 14, and 21. ZIC and SOX were the two most highly correlated genes in the anti-Hu patient data, with 7/13 patients with ZIC reactivity also enriching for a concurrent SOX peptide.

### **2.3.11 Novel PND/Cancer Associated Signatures in Anti-Yo Patients**

Beyond established PND and cancer-associated biomarkers, the data revealed several potentially novel disease-associated peptide binding signatures. After adjusting for multiple tests (Benjamini-Hochberg), PCM1 (Pericentriolar Material 1), SAP25

(Sin3A Associated Protein 25), ROBO4 (Roundabout Guidance Receptor 4), and MTMR14 (Myotubularin Related Protein 14) each show significant ( $p_{\text{adj}} \leq 0.05$ ) association with anti-Yo PND. Peptides derived from PCM1 showed the strongest association, with 24/36 anti-Yo patients and 0/50 healthy controls enriching a total of 7 unique peptides ( $p_{\text{Fisher's exact}} < 10^{-11}$ ). PCM-1 antibodies have been reported in cases of parainfectious acute cerebellar ataxia and autoimmune scleroderma,<sup>53,54</sup> but this is, to our knowledge, the first reported case of anti-PCM1 activity associated with Anti-Yo PND. Multiple peptides from SAP25 and ROBO4 were each identified in 15 patients (5 patients concurrently) while a single peptide from MTMR14 was common to 10 anti-Yo patients and zero healthy control sera.

## 2.4 DISCUSSION

We have designed and implemented the most complex and comprehensive human PhIP-Seq library to date, encompassing the entire human proteome and including all known and predicted splice variants and isoforms in the NCBI Protein database. In an effort to better understand autoimmune neurological diseases, we screened patient CSF and serum samples from two common paraneoplastic neurological disorders, the anti-Yo and anti-Hu syndromes. We chose these diseases for several reasons. First, while their autoantigens have been previously identified, the high-resolution epitope mapping afforded by PhIP-Seq or similar methods has, to our knowledge, not yet been achieved. Second, we reasoned that the high degree of alternative splicing in these neuronal antigens would help us leverage the unique design of our library which includes all published and computationally predicted splice variants.

Third, as these diseases are associated with a robust anti-tumor immune response, we reasoned that patient samples would likely enrich for multiple autoantigens that could have disease relevance and would be missed by more limited and hypothesis-driven antibody assays.

To our knowledge, the PhIP-Seq data from the anti-Yo patient cohort represent the highest-resolution epitope mapping of anti-Yo antigens to date. PhIP-Seq correctly identified the canonical anti-Yo antigens, CDR2 and CDR2L in 36/36 patients. This highly concordant result suggests that anti-Yo PND antibodies routinely bind linear epitopes without any requisite post-translational modifications (PTMs) or significant secondary or tertiary structure. In addition, CDR2L peptides were much more enriched than CDR2 peptides. Indeed, across all samples, phage expressing CDR2L peptides represented 21% of all phage sequenced, and these peptides were represented in the top 5 most significantly enriched sequences in > 80% of samples tested. While these data do not provide a quantitative measurement of antibody affinity (especially to the native form *in vivo*), they suggest that CD2RL is the immunodominant antigen in patients suffering from anti-Yo PND. This also supports previous studies arguing that the antibody responses to each protein are independent (i.e., not cross-reactive) and that CDR2L is the primary antigen.<sup>55</sup>

Within the anti-Hu cohort, PhIP-Seq yielded significant enrichment of nELAVL peptides in 17/34 patients. Those patients that were positive for nELAVL peptides yielded a remarkably convergent antigenic signature, with antibodies binding a short 17-residue sequence (QAQRFRLDNLNLMAYGVK) shared by ELAVL2, 3 and 4 but absent in ELAVL1. This short sequence lies at the junction of exons 6 and 7a in ELAVL4, and

high resolution deep mutational scanning followed by motif analysis led to a further refined signature motif sequence (RLDNLLNMAY) as most deterministic for antibody binding. We note that this short sequence has been previously characterized as a T cell epitope in anti-Hu patients.<sup>42,43</sup> Regarding the anti-Hu patients that did not yield peptides from nELAVL genes, it may be the case that the autoimmune antibodies in these patients may require specific secondary and tertiary conformations, somatic mutations, or PTMs not emulated by our peptide library.

Interestingly, our analysis of publicly available RNAseq datasets and our own amplicon sequencing of human thymic epithelial cells reveals that exon 7a, immediately adjacent to our identified motif, is absent in >99% of ELAVL4 transcripts in the thymus but abundant in specific neuronal subtypes and brain regions in mice. While investigation of exon exclusion in the thymus and the mechanisms underpinning autoimmunity are beyond the scope of this initial survey, we speculate that the absence of these short exons during the development of central tolerance could potentially influence the rise of autoreactive T lymphocytes when the full-length protein is encountered in the CNS or expressed by peripheral tumors. Further, exclusion of exons bearing proteasomal cleavage sites (as with exon 7a) could bias MHC processing and presentation. Lineages of B cells with B cell receptors specific for ELAVL4 isoforms containing these excluded exons could presumably be maintained and expanded through interactions with CD4<sup>+</sup> T cells specific for this short sequence when displayed on the surface of a B cell acting in its capacity as an antigen presenting cell. Such a collaborative process and mechanism is in keeping with previous studies examining autoimmune demyelinating disease and Nova2 PND.<sup>44,56</sup>

A major advantage of unbiased proteomic approaches like PhIP-Seq is the ability to interrogate patient samples for multiple antigens simultaneously, enabling high-throughput antigen-antigen and antigen-disease associations. By comparing the PND patient data with a collection of 50 healthy sera run on our platform, we identified several known and potentially novel antigens and/or biomarkers associated with each disease; most notably, ZIC and SOX family genes in the anti-Hu patients and PCM-1 and SAP25 in the anti-Yo cohort. The biological and clinical significance of these novel associations remain unknown. The clinical data on the specific cancer diagnosis and detailed neurologic phenotyping for each patient was not available for this study, and thus precludes analyses contrasting tumor type and particular neurologic deficits with the patient-level peptide binding signatures. Much like transcriptomic analyses, specific sets of correlated or biologically related antigens may provide researchers and clinicians with insight into yet undiscovered or previously unmeasurable cellular processes. We speculate, for example, that antibody binding signatures to endogenous, intracellular and normally inaccessible antigens reflect an immune process related to clearance of intracellular debris following apoptosis of tumor cells during the onco-immune response.<sup>57</sup> The preponderance of ZIC and SOX related peptides in our anti-Hu cohort and their high levels of expression in SCLC, for example, supports the hypothesis that patient antibodies are providing a snapshot of the interior of tumor cells.<sup>15</sup> The potential value of such biomarkers may only be realized after hundreds or thousands of patient samples have been run and correlated with higher-resolution disease phenotypic datasets that include long-term clinical outcomes and tumor transcriptional profiling.

This validation study confirms and extends the comprehensive nature of PhIP-Seq peptidome libraries for the purpose of identifying novel autoantigens in patient groups. Ultimately, as comprehensive, proteome-wide serological screens are more widely adopted and replace conventional candidate-based approaches, we expect the diagnostic, prognostic, and research potential of such assays to be fully realized. Indeed, the third most common category of autoantibody-mediated causes of encephalitis belongs to patients with unidentified autoantigens that have defied identification by more conventional methods.<sup>2</sup> Furthermore, high-resolution datasets characterizing the co-occurrence of specific antigens and antigenic determinants to specific diseases, symptoms, and clinical outcomes may ultimately serve to identify early biomarkers and may even stratify patient outcomes and inform treatment decisions. Finally, higher resolution epitope mapping and analysis of antigenic signatures across multiple patients should better inform the currently incomplete models describing onco-immunologic processes and the loss of self-tolerance underpinning the rise of PND and autoimmune disease in general.



## 2.5 MATERIALS AND METHODS

### 2.5.1 Computational Design of Library

As a basis for the library, all human sequences in the NCBI protein database (Nov 2015) including all splicing isoforms and computationally predicted coding regions were downloaded. Full length sequences were clustered on 99% sequence identity (CD-HIT v4.6)<sup>30</sup> to remove duplicate and partial sequences before computationally dividing them into 49 amino acid peptides using a 24 AA sliding window approach starting at the N-terminus. As such, each peptide shared a 25-residue overlap with the one preceding it on the primary protein sequence. Peptides encoding the final 49-residues at the C-terminus were substituted when the final sliding window resulted in shortened or truncated sequences (proteins not evenly divisible by 49). The resulting peptides were further clustered/collapsed on 95% identity such that peptides with 2 or fewer amino acid differences were combined by randomly choosing one representative sequence.

Amino acids sequences were converted to nucleotides using preferred *E. coli* codons (Table 2.2) and relevant restriction sites (EcoRI, HindIII, BamHI, XhoI) removed with synonymous mutations to facilitate cloning. To enable amplification from synthesized oligo pools, 21 nt universal priming sites were added to the 5' and 3' ends of each oligo. These sequences also serve to encode STREP and FLAG tag sequences (after addition of terminal codons by PCR) that allow for downstream assessment of proper, in frame cloning. The resulting library consists of 731,724 unique 189mer oligonucleotide sequences encoding 49AA human peptides flanked by 8AA STREP and FLAG sequences after amplification with appropriate primers.

## 2.5.2 Cloning into T7 Select vector

ssDNA oligonucleotides were commercially synthesized in three separate pools by Agilent Technologies (Santa Clara, CA). Oligos were amplified by PCR primers adding relevant EcoRI and HindIII restriction sites. 200 pg of DNA ( $10^8$  molecules per reaction across 50 reactions) was used as a template for each PCR reaction:

Reagent	Vol. ( $\mu$ l)	Concentration
Input DNA library	1	200 pg/ $\mu$ l
Platinum HiFi master mix	10	5X
Forward Primer* (adds EcoR1 site)	2	10 $\mu$ M
Reverse Primer* (adds HindIII site)	2	10 $\mu$ M
H <sub>2</sub> O	35	-

\*primer design:

forward: CTACGAATTCCTGGAGCCATCCGCAGTTCCG  
reverse: CTACAAGCTTCTTATCATCGTCGTCCTTG TAGTC

Thermocycling Conditions:

Temperature ( $^{\circ}$ C)	Time (mm:ss)	# cycles
98	02:00	1
98	00:30	25
68	00:30	
72	00:30	
72	05:00	1
4	Hold	-

PCR products were column cleaned (Zymo DNA clean and concentrator, Zymo Research, Irvine, CA) and eluted in 20  $\mu$ l H<sub>2</sub>O. DNA was then subject to restriction digestion under the following reaction conditions (incubated at 37 for 1 hour):

<b>Reagent</b>	<b>Vol. (μl)</b>	<b>Concentration</b>
Cleaned PCR product (DNA)	10	200 ng/μl
CutSmart buffer (NEB)	5	10X
EcoRI HF (NEB)	2	20,000 U/ml
HindIII HF (NEB)	2	20,000 U/ml
H <sub>2</sub> O	31	-

To minimize the proportion of truncations and deletion errors most commonly associated with commercial synthesis chemistries, digested fragments were size selected by agarose gel electrophoresis (Blue pippin 3% cassette (BDQ3010)) before ligation/cloning.

### 2.5.3 Cloning and Packaging

The digested oligonucleotides were cloned and packaged into competent phage following the T7 select cloning manual. The following reaction mix was incubated at room temperature for 15 minutes:

<b>Reagent</b>	<b>Vol. (μl)</b>	<b>Concentration</b>
Digested, size-selected DNA	1	6 ng/μl
T7 Vector Arms (EMD)	3	500 ng/μl
Quick Ligase Buffer (NEB)	5	2X
Quick Ligase (NEB)	1	10,000 U/ml

2 μl of the ligation reaction were added directly to 10 μl of T7 packaging extract (EMD) and allowed to incubate at room temperature for 2 hours. Reaction was quenched by adding 200 μl of chilled, sterile, LB. 24 separate 10 μl packaging reactions were carried out to ensure library complexity. Packaging efficiency was determined by plaque assay (as described in T7 Select Cloning Manual).

#### **2.5.4 In vivo amplification**

BLT5403 *E. coli* cultures were grown to log phase (OD<sub>600</sub> = 0.5), inoculated (0.001 MOI) with packaging reaction, and allowed to clarify (complete lysis) by incubating in 37 °C incubator (2-3 hours). BLT5403 carries an ampicillin-resistant plasmid expressing a wildtype gene 10A behind a T7 promoter. Phage produced in this strain carry 5-15 copies of the 10B capsid protein bearing the cloned fusion peptide. Lysates were cleared of debris by centrifugation at 4,000 rcf at 4 °C for 30 min before filtering through a 0.22 micron filter.

#### **2.5.5 Concentrating and Storing Phage Library**

Phage were precipitated by adding 1/5 volume 5x PEG/NaCl precipitation buffer (PEG-8000 20%, NaCl 2.5 M) and centrifuged at 13,000g for 60 minutes. Pellets were raised in 1/4 starting volume storage buffer (20 mM Tris-HCl, pH 7.5, 100 mM NaCl, 6mM MgCl<sub>2</sub>).<sup>58</sup> Stocks were titered by plaque assay and diluted/concentrated to ~10<sup>11</sup> pfu/ml.

#### **2.5.6 Immunoprecipitation**

All liquid handling steps were carried out on Biomek FX robotics platform (Beckman Coulter, Brea, CA). 96-well, full skirted, low profile PCR plates (BioRad Inc, Hercules, CA) were incubated overnight with 180 µl IP blocking buffer (3% BSA, PBS-T) to prevent nonspecific binding. Blocking buffer was replaced with 150 µl of phage library (10<sup>11</sup> pfu in storage buffer) and mixed with 2 µl patients CSF, patient sera diluted 1:50 in blocking buffer, or 2 ng commercial antibody and incubated O/N at 4°C.

Protein A and G magnetic beads (Dynabeads – Invitrogen, Carlsbad, CA) were mixed equally, washed three times by magnetic separation and resuspension in equivalent volumes of TNP-40 wash buffer (150 mM NaCl, 50 mM Tris-HCL, 0.1% NP-40, pH 7.5). 10 µl of A/G bead slurry was added to each IP reaction (using wide bore pipette tips) and incubated for 1 hour at 4°C. Bead-antibody complexes were washed three times by magnetic separation, removal of supernatant and resuspension in 150 µl TNP-40 wash buffer. Subsequent to final wash, beads were re-suspended in 150 µl chilled LB and used to inoculate fresh *E. coli* cultures (400 µl at OD600 = 0.5) for *in vivo* amplification. Lysates were clarified by centrifugation at 3,000g (at 4 degrees) and 150 µl removed/stored for NGS library prep or additional rounds of IP.

### 2.5.7 NGS Library Prep

Sequencing libraries were prepared from lysates with a single PCR reaction using multiplexing (MP) primers as follows:

Reagent	Vol. (µl)	Concentration
<i>E. coli</i> lysate	2	Diluted 1:50
Phusion HF Buffer (NEB)	10	5X
Forward MP Primer*	2.5	10 µM
Reverse MP Primer*	2.5	10 µM
dNTPs	1	10 mM
Phusion Polymerase	0.5	1U/µl
H <sub>2</sub> O	31.5	-

**\*Multiplexing Primer Design:**

Forward:

AATGATACGGCGACCACCGAGATCTACAC[NNNNNNN]GGAGCTGTCGTATTCCAG  
TCAGGTGTGATGCTC

NNNNNNN = i5 index

Reverse:

CAAGCAGAAGACGGCATAACGAGAT[NNNNNNN]GGTAACTAGTTACTCGAGTGCG  
GCCGCAAGC

NNNNNNN = i7 index

Thermocycling conditions:

Temperature (°C)	Time (mm:ss)	# cycles
95	02:00	1
95	00:30	5
63	00:30	
72	01:00	
95	00:30	5
60	00:30	
72	01:00	
95	00:30	10
58	00:30	
72	01:00	
72	02:00	1
4	hold	-

Libraries were cleaned and size selected by Ampure XP magnetic beads (0.8X volume, 40 µl) (Beckman Coulter, Pasadena, CA), eluted in 20 µl of nuclease-free water and quantified by Qubit (Thermo Fisher, Waltham MA) dsDNA high sensitivity fluorimeter. Sequencing was performed on either an Illumina HiSeq 2500 or 4000 using custom indexing and sequencing primers:

Read 1 sequencing primer: TCCTGGAGCCATCCGCAGTTCGAGAAA

Read 2 sequencing primer: AAGCTTCTTATCATCGTCGTCCTTGAGTC

i7 indexing primer: GGCCGCACTCGAGTAACTAGTTAACC

i5 indexing primer: CACCTGACTGGAATACGACAGCTCC

### **2.5.8 Bioinformatic Analysis**

Reads were quality filtered, paired-end reconciled (PEAR v0.9.8),<sup>59</sup> and aligned to a reference database of the full library (bowtie2 v2.3.1).<sup>60</sup> Sam files were parsed using a suite of in-house analysis tools (Python/Pandas) and individual phage counts were normalized to reads per 100k (rpK) by dividing by the sum of counts and multiplying by 100,000. Peptide-level fold changes were calculated by dividing individual peptide rpK values with their expected abundance (mean rpK in mock IPs). For peptides never observed in any control IP, fold changes were calculated using the median rpK for all peptides derived from that gene in the mock IP.

We identified in our control IP's a set of the 100 most abundant peptides that also have a standard deviation less than their mean. This "internal control" set represented the most abundant and consistent phage carried along specifically by the protein-AG beads or other reagents, in absence of any antibody. We calculate a sample-specific scaling factor, defined as the ratio of median abundances (rpK) of these 100 peptides in our controls to their abundance in the given sample. Fold change values are the multiplied by their sample-specific scaling factor. These top 100 peptide sequences are available at <https://github.com/derisilab-ucsf/PhIP-PND-2018>.

To account for the bias introduced in the fold-change calculation stemming from under-represented peptides in the mock IP having inflated fold-changes despite only modest representation in an experimental IP, we normalize, within each experimental batch (defined as all samples on a given 96-well plate - never less than 50 samples), fold-change values using a correction factor defined as the solution to a linear regression fit to the fold-change values as a function of their mean  $\log_2(\text{rpK})$  in AG-only controls (Figure 2.8). Importantly, fold-change values are only corrected when the correction factor is positive (expected fold-change given the abundance in mock IPs is positive). We fit a normal distribution to the scaled and normalized values and assign p-values by evaluating the survival function (1-CDF) at a given fold-change (Figure 2.8). To account for multiple tests, we calculate an adjusted p-value ( $p_{\text{adj}}$ ) using the Benjamini-Hotchberg procedure, enforcing a false discovery rate (FDR) of 0.05. Peptides were reported as significant as those with a positive fold-change with a  $p_{\text{adj}} \leq 0.05$  in both replicates.

### **2.5.9 Design of Hu Motif Mutational Scanning Library**

Degenerate 112 nt oligos were designed encoding 24 amino acids (PLHHQAQRFRLDNLNLMAYGVKRLMKL) spanning the 17-residue motif (QAQRFRLDNLNLMAYGVK) identified in the PhIP-Seq data. A unique oligo containing a degenerate NNN codon at each position was ordered/synthesized (IDT, San Diego, CA). Sequences were flanked with relevant restriction sites (EcoRI, HindIII) to enable in-frame cloning. Oligos were pooled in equimolar ratios and cloned following the manufacturer's protocol in the T7 Select Cloning Manual. Phage libraries were



subjected to 3 rounds of immunoprecipitation using patient antibodies (or AG beads alone) following the standard PhIP-Seq protocol described above. Final phage populations were sequenced with an average of ~2M 150 base paired-end reads on an Illumina MiSeq. Reads were quality filtered, reconciled/merged (PEAR),<sup>59</sup> and translated to amino acid sequences. At each position, the antibody binding preference was defined as the  $\log_2$ (fold-change) of each amino acid substitution versus AG-only controls and relative to the endogenous/reference residue, quantified accordingly:

Binding preference =

$$(\log_2(P_{aa,j})_{Ab} - \log_2(P_{aa,j})_{AG}) - (\log_2(P_{wt,j})_{Ab} - \log_2(P_{wt,j})_{AG})$$

where:

$P_{aa,j}$  = proportion of reads with specific amino acid substitution (aa) at position j

$P_{wt,j}$  = proportion of reads with wildtype sequence (wt) at position j

The mutational tolerance was then defined as the mean binding preference at each position.

#### **2.5.10 ELAVL4 amplicon sequencing from bulk thymus and human TECs**

Sorted TECs were isolated from human thymus obtained from 18- to 22-gestational-week specimens under the guidelines of the University of California San Francisco Committee on Human Research. Tissue was washed, cut into small pieces using scissors, and gently mashed using the back of a syringe to extract thymocytes. To isolate TECs, remaining tissue pieces were digested at 37 °C using medium containing

100 µg/ml DNase I (Roche, Belmont, CA) and 100 µg/ml Liberase™ (Sigma-Aldrich, St. Louis, MO) in RPMI. Fragments were triturated through a 5-ml pipette after 6 and 12 min to mechanically aid digestion. At 12 min, tubes were spun briefly to pellet undigested fragments and the supernatant was discarded. Fresh digestion medium was added to remaining fragments and the digestion was repeated using a glass Pasteur pipette for trituration. Supernatant from this second round of digestion was also discarded. A third round of enzymatic digestion was performed using digestion medium supplemented with trypsin–EDTA for a final concentration of 0.05%. Remaining thymic fragments were digested for another 30 min or until a single cell suspension was obtained. The cells were moved to cold MACS buffer (0.5% BSA, 2 mM EDTA in PBS) to stop the enzymatic digestion. Following digestion, TECs were enriched by density centrifugation over a three-layer Percoll gradient with specific gravities of 1.115, 1.065 and 1.0. Stromal cells isolated from the Percoll-light fraction (between the 1.065 and 1.0 layers) were washed in MACS buffer. For surface staining, cells were blocked using a human Fc Receptor Binding Inhibitor Antibody (eBioscience, San Diego, CA) and incubated on ice for 30 min with the following antibodies:

PE anti-human Epcam (HEA-125) (Miltenyi 130-091-253) - dilution 1:50

Alexa Fluor 488 anti-human CD45 (HI30) (BioLegend 304017) - dilution 1:100

Epcam+ CD45- DAPI- cells were sorted using a FACS Ariall (BD Biosciences, San Jose, CA) directly into TRIzol LS (ThermoFisher Scientific, South San Francisco, CA). The aqueous phase was extracted using the manufacturer's instructions followed

by RNA isolation using the RNeasy micro kit (QIAGEN, Redwood City, CA). RNA was reverse transcribed using iScript cDNA synthesis kit (Bio-Rad, Emeryville, CA).

RNA samples from whole fetal and adult human thymus were purchased from Agilent (Santa Clara, CA) and Clontech (Mountain View, CA). RNA was reverse transcribed using iScript cDNA synthesis kit (Bio-Rad, Emeryville, CA).

Amplicon libraries spanning the relevant exons were generated using the following primers:

ELAVL4\_fwd: GAACCGATTACTGTGAAGTTTGCCAAC

ELAVL4\_rev: GTAGACAAAGATGCACCACCCAGTTC

PCR products were purified using the QIAGEN Gel Extraction kit and library construction was conducted on the quantified DNA using the NEBNext® Ultra™ DNA Library Prep Kit (New England Biolabs, Ipswich, MA). The final library was sequenced using a MiSeq with single-end, 300-base reads.

### **2.5.11 Motif Discovery**

The redundant and overlapping nature of our library design presents some potentially confounding issues with existing motif discovery software packages leveraging hidden Markov models (HMMs) to identify significantly represented subsequences and motifs in a collection of sequences or strings. Highly redundant and replicate sequences as an input dataset confound or overwhelm the motif finding algorithms, often resulting in the most significant motifs reported as those in the

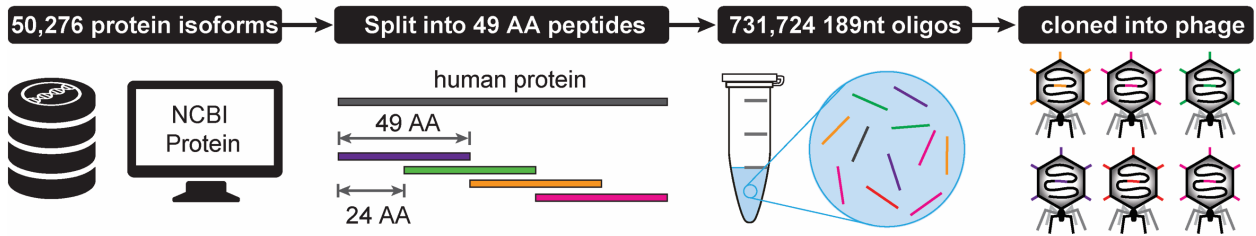
overlapping regions of adjacent peptides. To counter this, we manually curated patient results to eliminate sequences with >50% identity. We then iteratively searched each set of results for peptides with 10 or more identical sequential residues and chose as a representative sequence, that with the higher enrichment score (lowest adjusted p-value). These curated datasets were examined by the researchers before being analyzed for motifs by GLAM2, part of the meme-suite software package.<sup>61</sup>

#### **2.5.12 Clinical PND Antibody Confirmation**

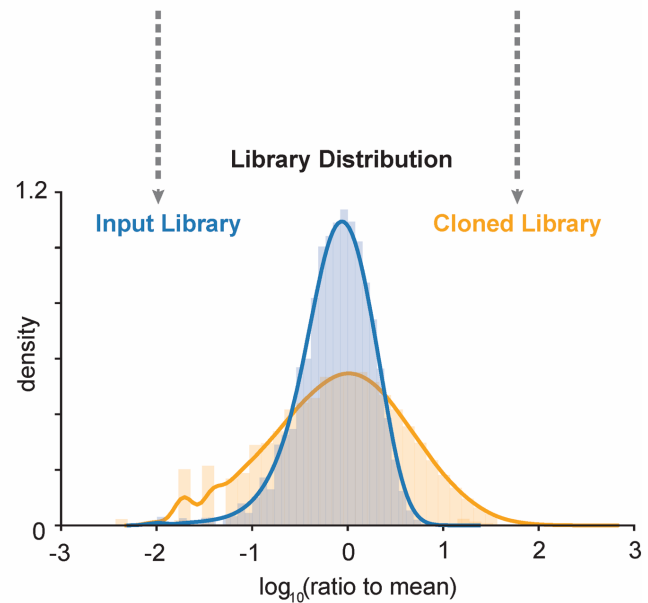
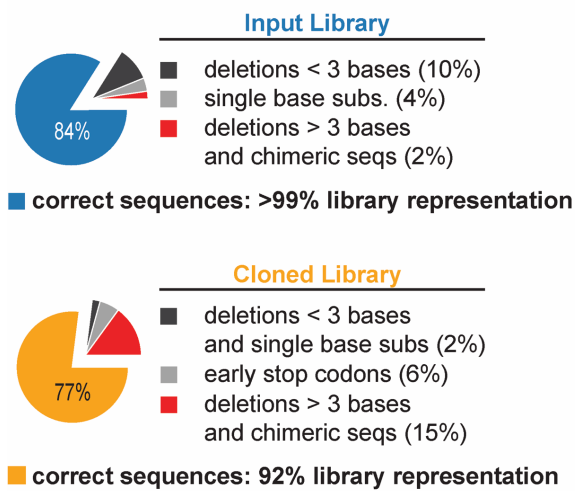
Confirmation of anti-Yo and anti-Hu activity in patient samples was performed in the Mayo Clinic Neuroimmunology Laboratory with CLIA and New York State approved methodology. Briefly, patient samples (serum/CSF) were tested on a mouse composite tissue slide with an indirect immunofluorescence assay for IgG binding corresponding to the anti-Hu or anti-Yo pattern, as previously described.<sup>39</sup> The results were confirmed with western blot analysis on rat cerebellum preparations.

## 2.6 FIGURES

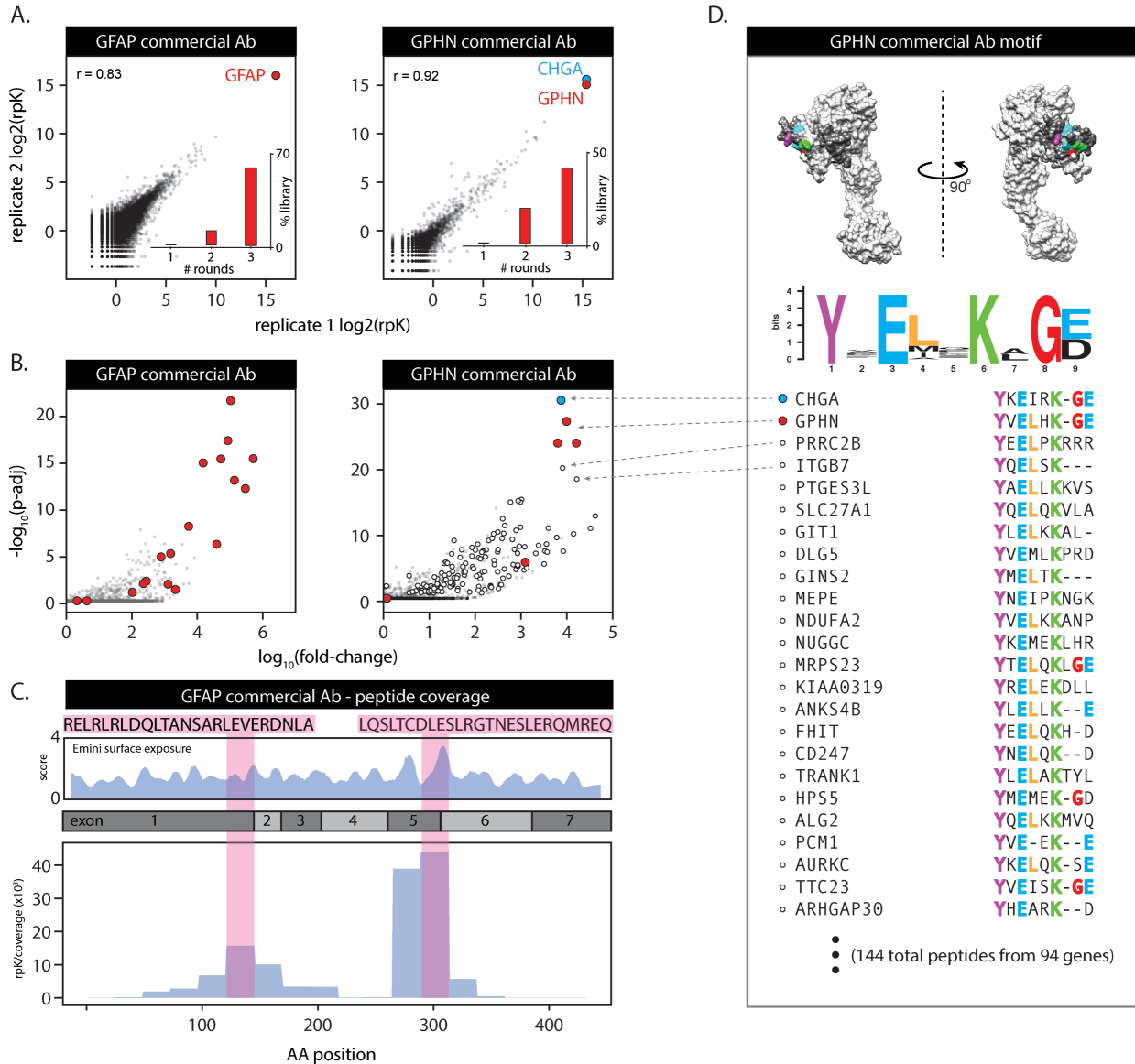
### A. Library Design



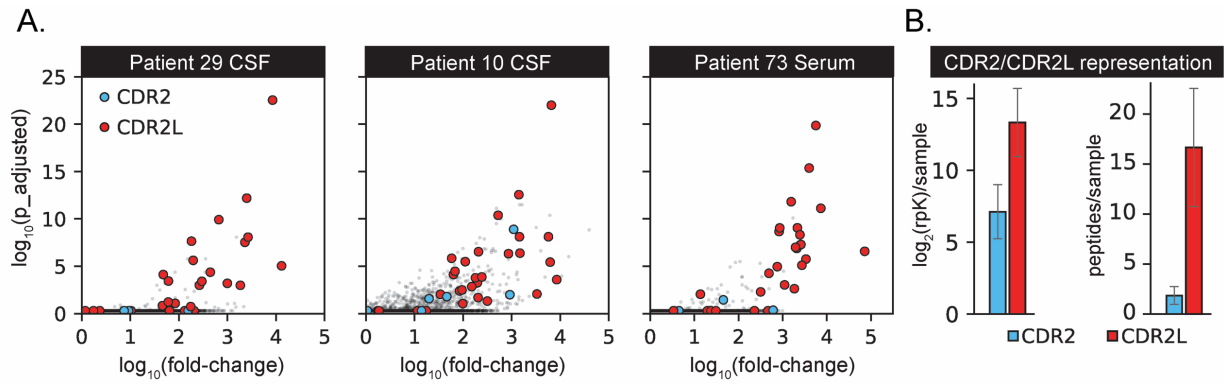
### B. Library QC



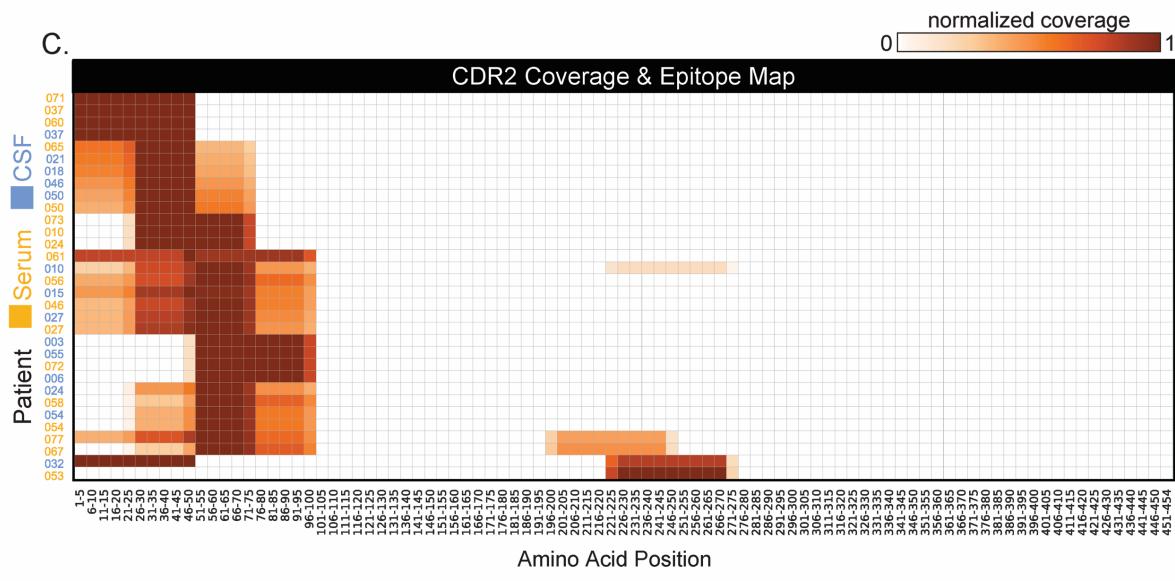
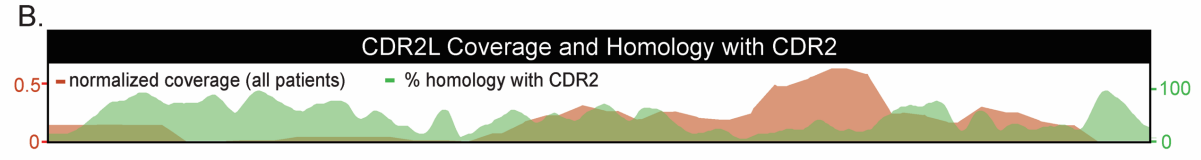
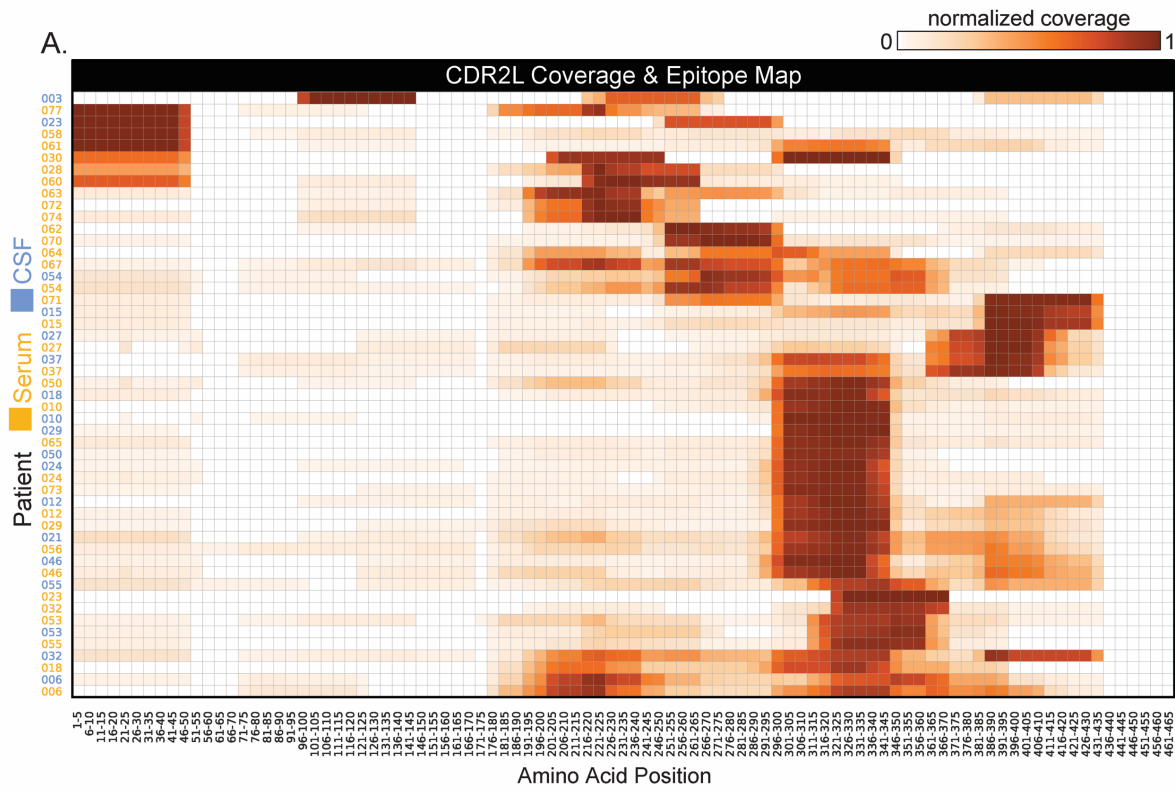
**Figure 2.1 - Design and Characterization of Library.** A.) All human protein isoforms, variants, and computationally predicted coding regions were downloaded from the NCBI Protein database. Full length sequences were clustered on 99% sequence identity (CD-HIT v4.6) to remove duplicate and partial sequences. Each protein was computationally divided into 49 amino acid peptides using a 24 AA sliding window. Redundant or duplicate sequences were removed by further clustering on 95% sequence identity. The resulting 731,724 sequences were synthesized and cloned into the T7 select vector (Millipore, Burlington, MA). B.) Assessment of the library quality before (blue) and after (yellow) cloning and packaging.



**Figure 2.2 - Validation of library by commercial antibody IP's.** Phage library was subjected to three rounds of selection by polyclonal commercial antibodies to GFAP and GPHN. A.) Replicates show high correlation of gene-level counts for IPs performed on separate days. Final libraries are dominated by the commercial antibody target. The GPHN antibody also consistently enriched a single peptide from chromogranin A (CHGA). B.) Scatterplots of peptide-level enrichments show 18 unique GFAP peptides and 4 GPHN peptides were identified as antibody binders (red markers). Peptides sharing a motif with both GPHN and CHGA (white, see D). C.) Alignment of GFAP peptides to the full-length protein reveals two distinct regions of antibody affinity corresponding to regions of high solvent exposure and B cell antigenicity as predicted by IEDB Tools Emini surface exposure and Bepipred algorithms. D.) Motif analysis of the significant peptides in the GPHN IP reveal a short, discontinuous motif shared by 144 significantly enriched peptides in the final population (from 94 unique genes/proteins), including the most abundant CHGA peptide. The motif is highlighted on the crystal structure of GPHN (top) and peptides sharing this motif are marked by a white circle in B.

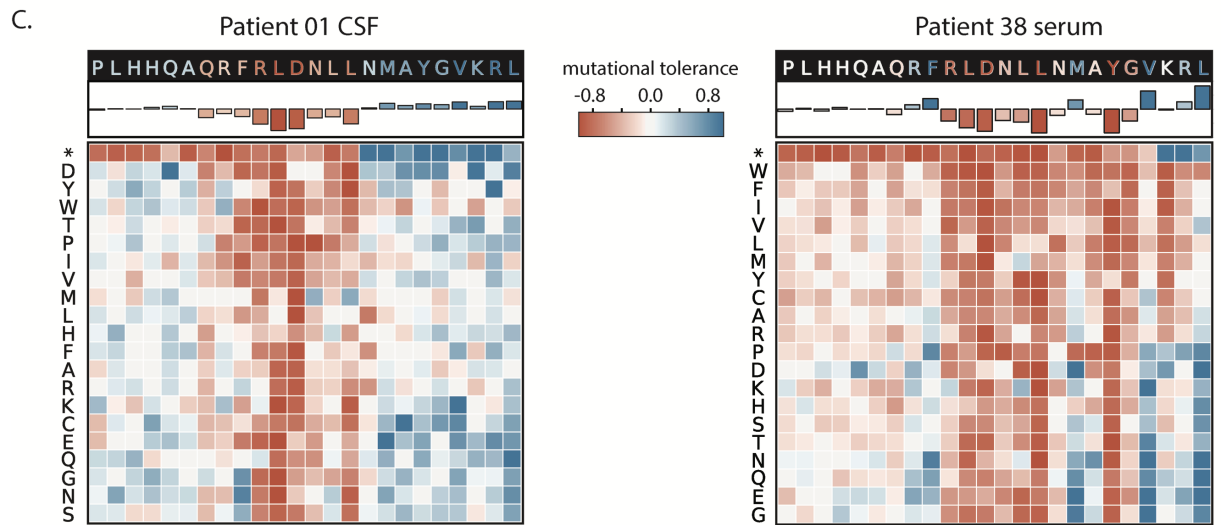
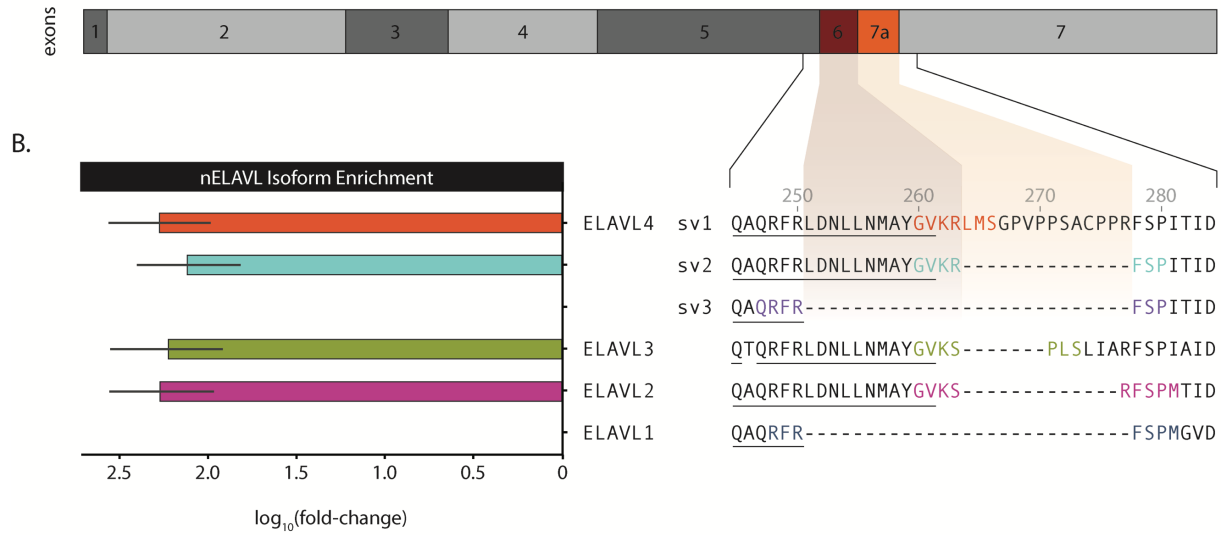
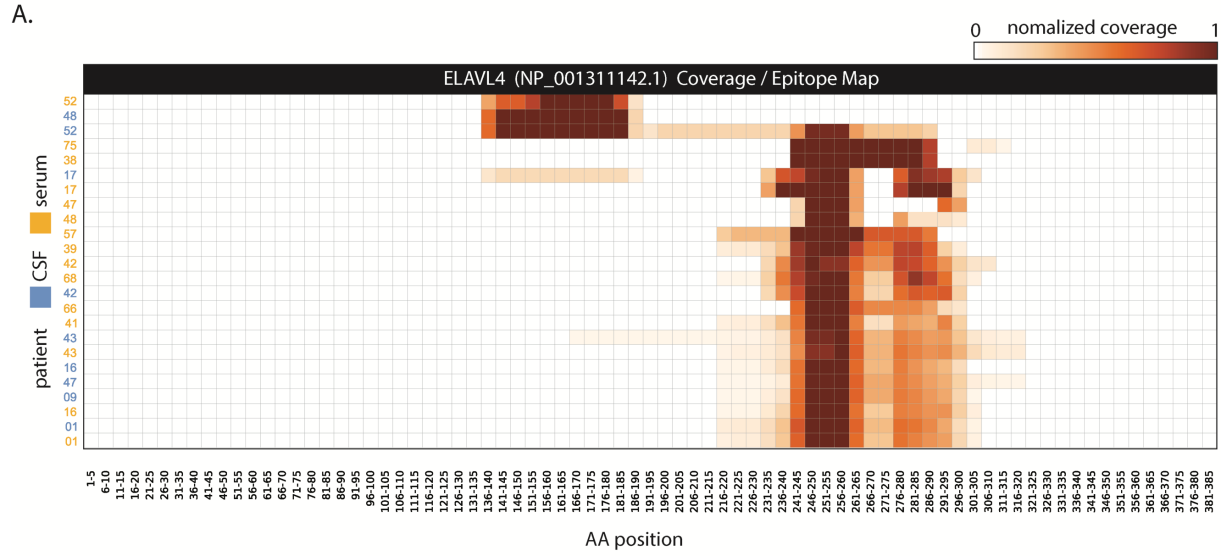


**Figure 2.3 - Anti-Yo representative results.** A.) Representative data from 3 patients demonstrating the robust enrichment of CDR2/CDR2L peptides. CDR2/CDR2L peptides were the top 5 most abundant peptides in 44/51 samples. B.) While CDR2/CDR2L peptides represented 21% of all phage, CDR2L enrichment is more pronounced and complex (more unique peptides per sample) than CDR2.

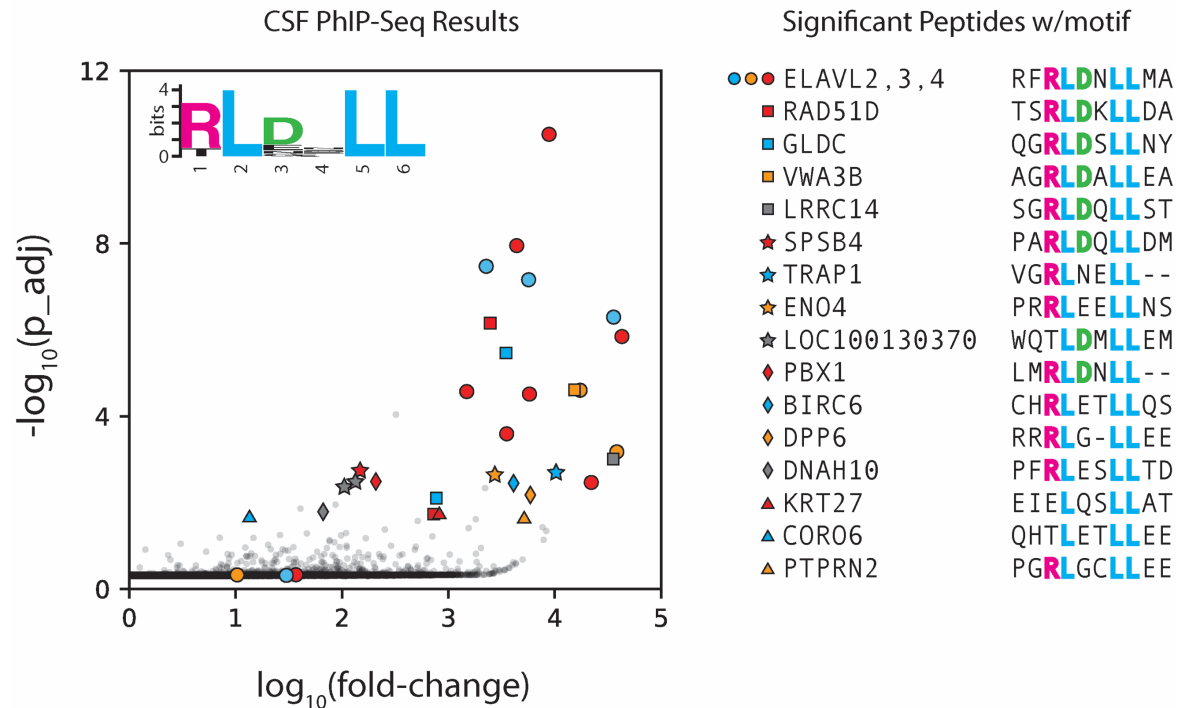




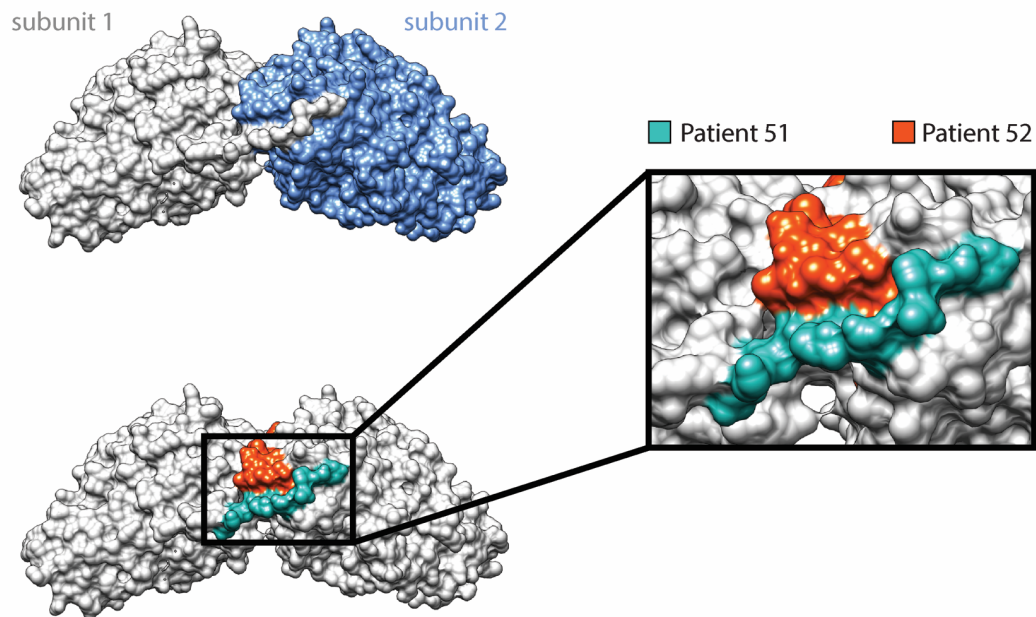
**Figure 2.4 - Anti-Yo epitope mapping.** Normalized coverage of each samples' significant CDR2L peptides aligned to the full-length gene (divided into 5 amino acid bins). Patient binding signatures vary, though a majority converge on residues 322-346. B.) Patient antibodies targeting CDR2L bind to regions of least homology with CDR2, suggesting the antibody responses to each protein are independent. C.) Epitope map showing normalized peptide coverage across the full length CDR2 sequence. Patient antibodies enrich for peptides primarily restricted to the first 100 amino acids at the N terminus.



**Figure 2.5 - Anti-Hu epitope mapping.** A.) Antibody binding signatures for samples that enriched nELAVL peptides. A majority of patients show a markedly convergent binding preference for a short 17-residue sequence at the exon 6/7a junction of ELAVL4. B.) Mean fold-change values for peptides spanning the unique regions defining ELAV4 splice variants and those from ELAVL2 and ELAVL3, indicating lack of enrichment beyond the exon 6/7a junction. C.) Mutational scan using smaller phage library with all possible point mutations at each location spanning the motif highlighted in (B). Mutational tolerance is calculated for each position, defined as the fold-change over unselected (AG-bead only IP) relative to the reference sequence. Most telling are the mutations encoding stop codons (denoted \*) that reveal patient 01's antibodies require a minimal sequence up to residues 255 (RLDDNLL) with the subsequence QRFRLDNLL least amenable to mutation. Patient 38's antibodies bind to residues up to and including position 260 (RLDNLLNMAYGV), with highest affinity for RLDNLLN-AYG.

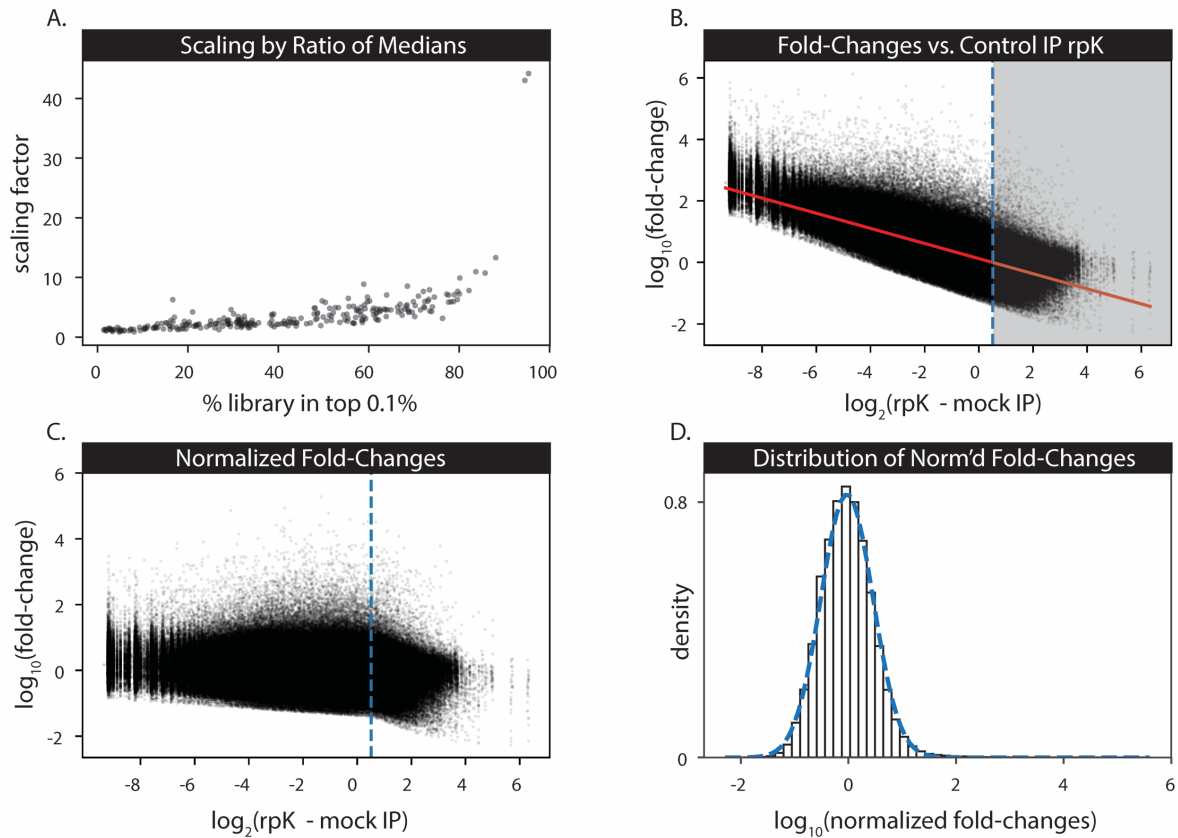


**Figure 2.6 - Patient 01 samples dominated by ELAV motif.** Immunodominant nELAVL motif in Patient 01 CSF sample is that identified by mutational scanning. The short sequence, dominated by RLDxLL is present in 34 of the 36 peptides identified in both replicates. Apart from nELAVL proteins, the patient's antibodies enrich for peptides from 16 additional genes containing the motif. The biological consequences of such promiscuous autoantibody binding are unknown.



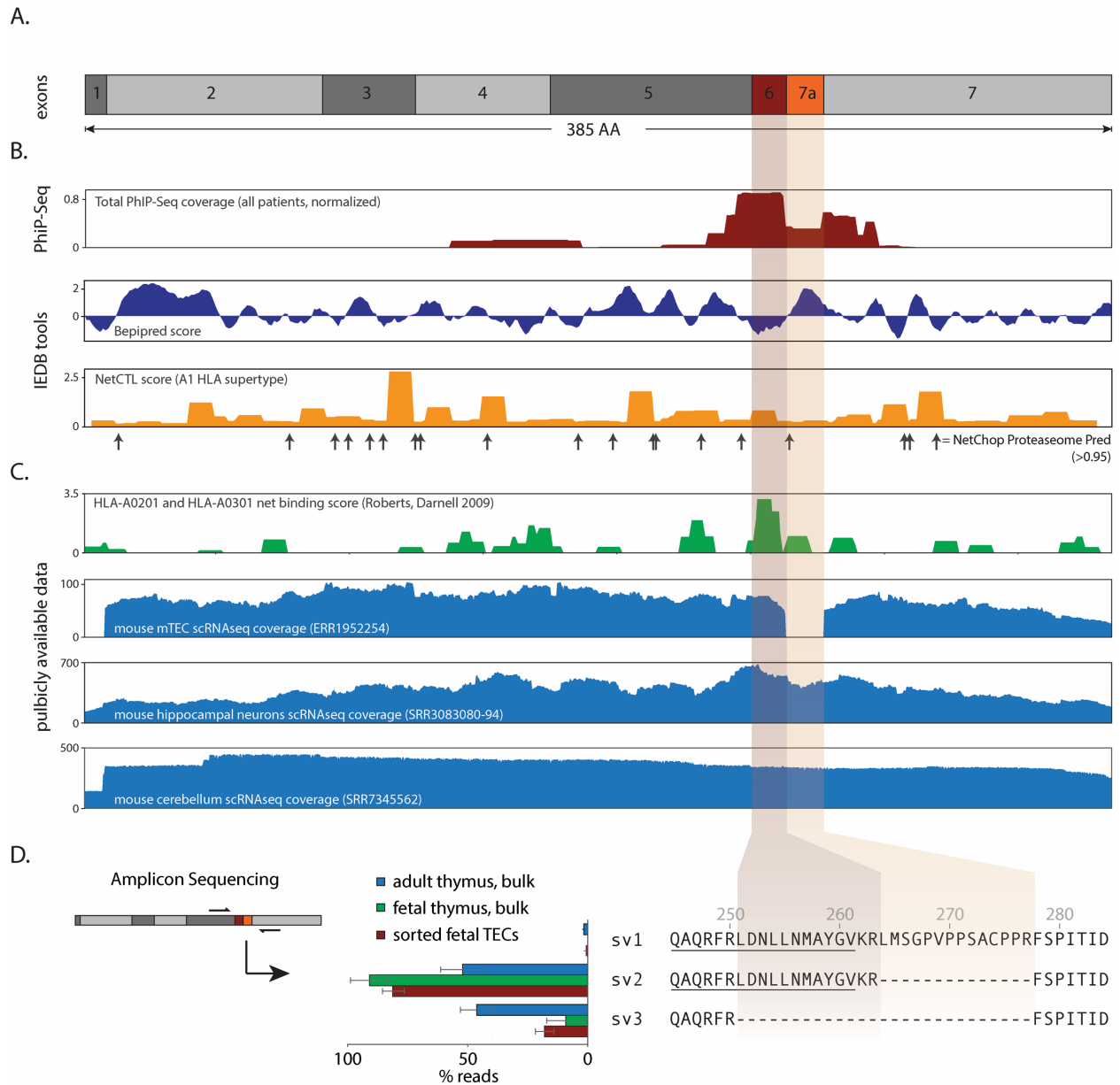
**Figure 2.7 - Anti-CRMP5 epitope mapping.** Two anti-Hu PND patient (patients 51 and 52) CSF samples also enriched CRMP5 peptides. CRMP5 is a well-established clinical biomarker for SCLC and malignant thymoma and anti-CRMP5 antibodies are often found concurrently with anti-Hu antibodies in patients with a PND. Though the peptides identified in each patient were discontinuous on the primary amino acid sequence, they are in close proximity in the 3D protein structure of CRMP5 at the region surrounding the C-terminus and the interface of the CRMP5 homodimer.

## Scaling and Normalizing Phage IP Data



**Figure 2.8 - Scaling and normalizing peptide fold-change values.** A scaling factor, defined as the ratio of median rpK values of the top 100 mock IP AG bead-binding phage (see Methods), was multiplied by the normalized peptide count for a given experimental sample. A.) Plotting the scaling factor vs the proportion of an experimental sample's final phage population present at or above the top 0.1% of rpK shows that samples dominated by a phage scale more dramatically. B.) Linear regression of scaled fold-change values to peptide representation in mock IPs. Intuitively, peptides with low abundance in the control IPs have higher calculated fold changes despite low rpK in the final population. Fold-change values were normalized by subtracting the solution to the fitted line when the expected fold-change was greater than 0. The resulting shifted/normalized fold changes (C.) are centered around 0 and normally distributed (D).

## ELAVL4 Exon Inclusion/Exclusion



**Figure 2.9 - PhIP-Seq Identifies short exon in ELAVL4 excluded in thymus.** A.) Exon map of ELAVL4. B.) Normalized PhIP-Seq coverage from all positive patients shows preferential binding of antibodies to exon 6. Exon 6 has low predicted B cell antigenicity but high prediction of MHC processing and presentation by IEDB tools Bepipred algorithm.<sup>62</sup> C.) Exon 6 contains top scoring MHC-I binding prediction score defined by Roberts, et al.<sup>43</sup> Analysis of publicly available datasets reveals exon 7a is excluded from transcripts in murine mTECs but included in mouse hippocampal and cerebellar neuron scRNAseq datasets. D.) Amplicon sequencing of human TEC cDNA corroborates exclusion of exon 7a in the thymus.

## 2.7 TABLES

PND	Antigen	Unique Peptides		# samples		$p_{\text{fisher's exact}}$	$p_{\text{adj}}$	Relevant Clinical Associations
		PND	Healthy	PND	Healthy			
Hu	SOX family	28	1	19	2	$4.0 \times 10^{-06}$	0.01	NSCLC, SCLC, <sup>51</sup> esophageal and bladder cancer, <sup>63</sup>
	ZIC family	25	0	13	0	$1.7 \times 10^{-05}$	0.01	SCLC, <sup>64</sup> ovarian and breast cancer, <sup>65 66</sup>
Yo	PCM1	7	0	24	0	$2.1 \times 10^{-12}$	$1.5 \times 10^{-9}$	Parainfectious ataxia, <sup>53</sup> scleroderma, <sup>54</sup> glioblastoma <sup>67</sup>
	SAP25	16	0	15	0	$3.6 \times 10^{-07}$	$2.0 \times 10^{-4}$	Ovarian cancer, <sup>68</sup> promyelocytic leukemia <sup>69</sup>
	ROBO4	4	1	15	1	$4.1 \times 10^{-06}$	$1.8 \times 10^{-3}$	Breast cancer, <sup>70</sup> bladder cancer, <sup>71</sup> glioma <sup>72</sup>
	MTMR14	1	0	10	0	$8.7 \times 10^{-05}$	$3.2 \times 10^{-2}$	none discovered

**Table 2.1 - Significant antigen-disease associations.** Comparison of PhIP-Seq patient binding data to a collection of 50 healthy serum samples reveals known and potentially novel PND-binding signatures. Genes showing significant associations after correcting for multiple tests ( $p_{\text{adj}}$ , Benjamini Hotchberg). SOX and ZIC family transcription factors are well established SCLC cancer biomarkers, and these antibodies are often associated with PND, particularly anti-Hu. In the anti-Yo cohort, PhIP-Seq enriched for PCM1, SAP25, ROBO4, and MTMR14.



<b>Amino Acid</b>	<b>Codon</b>	<b>Amino Acid</b>	<b>Codon</b>
A	GCG	L	CTG
C	TGC	N	AAC
B	GAT	Q	CAG
E	GAA	P	CCG
D	GAT	S	AGC
G	GGC	R	CGT
F	TTT	U	TGC
I	ATT	T	ACC
H	CAT	W	TGG
K	AAA	V	GTG
J	CTG	Y	TAT
M	ATG	X	GCT
		Z	GAA

**Table 2.2 - Preferred *E. coli* codons used in library design.**

## Chapter 3

# Metagenomic next-generation sequencing for chronic meningitis: a case series

### 3.1 ABSTRACT

Identifying infectious causes of subacute and chronic meningitis can be challenging. Enhanced, unbiased diagnostic approaches are needed. Here, we assess the utility of metagenomic deep sequencing (MDS) of cerebrospinal fluid (CSF) in diagnostically challenging cases of subacute and chronic meningitis. We present a case series of patients in whom metagenomic next-generation sequencing (mNGS), supported by a statistical framework leveraging mNGS sequencing data from non-infectious patients and environmental controls, successfully identifies the responsible pathogen. mNGS identified parasitic worms, fungi and viruses in seven subjects: *Taenia solium* (n=2), *Cryptococcus neoformans*, human immunodeficiency virus-1, *Aspergillus oryzae*, *Histoplasma capsulatum*, and *Candida dubliniensis*. Evaluating mNGS data with a weighted z-score based scoring algorithm effectively separated *bona fide* pathogen sequences from spurious environmental sequences. Prioritizing metagenomic data findings with a scoring algorithm greatly clarified data interpretation and highlights the

difficulties attributing biological significance to organisms present in control samples used for metagenomic sequencing studies.

### 3.2 INTRODUCTION

Subacute and chronic meningitis is diagnostically challenging given the wide range of potential infectious, autoimmune, neoplastic, paraneoplastic, parameningeal and toxic causes.<sup>73,74</sup> Securing a final diagnosis can require weeks or months of testing or remain unsolved, necessitating empiric treatment approaches that may be ineffective or even harmful.

Unlike traditional testing for specific microbes or categories of infection, metagenomic and metatranscriptomic next-generation sequencing (mNGS) of cerebrospinal fluid (CSF) or brain tissue screens for *all* potential CNS pathogens, and can identify novel or unexpected pathogens.<sup>75,76,77,78,79,80,81,82</sup> Multiple computational algorithms and pipelines have been developed to identify microbial sequences in mNGS datasets rapidly.<sup>83,84,85</sup> However, mNGS data requires careful analysis to determine which, if any, of the identified microbes represent a true pathogen versus environmental contamination, and failure to make this distinction has resulted in spurious disease associations with organisms later determined to be laboratory contaminants.<sup>86,87,88</sup>

This chapter describes a straightforward statistical approach to analyze mNGS data leveraging an extensive database of water-only controls (n=24) and surplus CSF samples (n=94) obtained from patients with clinically-adjudicated non-infectious neurologic diagnoses including autoimmune, neoplastic, structural and neurodegenerative etiologies (“control cohort”). This statistical approach quantifies the

uniqueness of observing a particular microbe in a patient sample at a given level of abundance (at multiple taxonomic levels) by comparison to its mean level of abundance across the control cohort. Here we report the utility of this statistical framework for identifying microbial pathogens in seven challenging cases of subacute or chronic meningitis, as well as for analyzing publicly-available data from recent mNGS infectious diagnostic and brain microbiota studies.<sup>89,10,90</sup>

### 3.3 RESULTS

The seven study subjects ranged in age from 10 to 55 years old, and three (43%) were female. Additional clinical details are listed in Table 3.1. In each case, the causative pathogen was identified within the top two scoring microbes identified by our algorithm (Figure 3.2). Across the seven study subjects, the mean of reported taxa was reduced by 87% (range 41-99%) if taxa with a combined score of zero or less were removed (i.e., average of 307 (range 11-1,313) taxa before filtering and 53 (range 1-297 after filtering)).

#### 3.3.1 Case Descriptions

##### *Taenia solium*

Participant 1 was a 29 year-old man from Nicaragua with headache and diplopia. CSF examination revealed an opening pressure of >50 cm, 66 WBC/mm<sup>3</sup> (89% lymphocytes, 4% neutrophils, 4% monocytes and 3% eosinophils), 1 RBC/mm<sup>3</sup>, total protein 43 mg/dL (normal range 15-50 mg/dL), and glucose 27 mg/dL (normal range 40-70 mg/dL). Contrast-enhanced brain magnetic resonance imaging (MRI) revealed

enhancement of the basilar meninges and several cranial nerves (Figure 2.3A). Although serum cysticercosis antibody was positive, there were no cysts or calcifications on brain MRI. The low CSF glucose, basilar meningitis, positive tuberculin skin test, positive tuberculosis interferon-gamma release assay, and his high-risk country of origin prompted empiric treatment for *Mycobacterium tuberculosis* (TB) meningitis. He improved clinically over the next three weeks but then worsened, requiring multiple lumbar punctures (LPs), MRIs and hospitalizations over the next year. Subsequent CSF samples showed worsening lymphocytic pleocytosis, persistently low glucose and elevated protein. Neuroimaging demonstrated persistent basilar meningitis and development of communicating hydrocephalus, again without cysts. His incomplete clinical response was initially attributed to treatment noncompliance, but after he worsened despite directly observed TB therapy, multi-drug resistant TB was suspected. After his fourth clinical decline (including discussion of ventriculoperitoneal shunt placement for worsening hydrocephalus) he was readmitted to the hospital for a new diagnostic work-up. Empiric therapy was broadened to include anti-helminthic treatment. CSF was submitted for research-based mNGS. The mNGS data demonstrated no sequences aligning to mycobacterial species. However, 58,789 unique, non-human read pairs aligned to the genus *Taenia* (Table 3.2), with the vast majority specifically mapping to the *Taenia solium* genome (Figure 3.2). A CSF specimen that had been sent concurrently for fungal 18s rRNA polymerase chain reaction (PCR) unexpectedly amplified the *Taenia solium* 18s rRNA gene, and CSF cysticercosis IgG antibody was positive. A FIESTA (Fast Imaging Employing Steady-state Acquisition) sequence protocol was included for the first time on the brain MRI,

and this revealed tangled hypointensities in the basilar cisterns, most prominently involving the prepontine cisterns with extension into the right internal auditory canal. After eight months of anti-helminthic treatment, he clinically improved to baseline except for a residual action tremor.

Participant 2 was a 34 year-old woman who had immigrated to the United States from El Salvador 13 years prior. She presented with nine months of right-sided headache, left facial numbness, right pulsatile tinnitus and recurrent loss of consciousness. Brain MRI showed hydrocephalus and right anterior temporal lobe and pre-pontine cysts (Figure 3.3B). CSF examination revealed an opening pressure of 36 cm, 115 WBC/mm<sup>3</sup> (1% neutrophils, 31% monocytes, 46% lymphocytes, 21% plasmacytoid lymphocytes and 1% eosinophils), 2 RBC/mm<sup>3</sup>, glucose <10 mg/dL and total protein 89 mg/dL. After her CSF and serum cysticercosis IgG antibody returned positive, she was treated with albendazole and prednisone for more than one month. However, she developed worsening neck pain, and a repeat CSF exam showed elevated intracranial pressure, pleocytosis with a new eosinophilia, undetectable glucose and elevated protein, raising concern for an alternative diagnosis. Her CSF mNGS data contained 569 read-pairs (Table 3.2) aligning to the genus *Taenia* (Figure 3.2). She was treated with dual anti-helminthic therapy with adjunctive glucocorticoids and had an excellent clinical response.

### *Cryptococcus neoformans*

Participant 3 was a 52 year-old man with a history of migraine and HIV-1 infection diagnosed in 2013 (viral load detectable but <40 copies/mL; CD4 count 20

cells/mm<sup>3</sup>). He also had a history of injection drug use (IDU), hepatitis C virus infection, *Staphylococcus aureus* endocarditis, and syphilis. He presented with agitation, confusion and ataxia. Because of his prior *S. aureus* bacteremia, syphilis, history of migraine and immunosuppressed state, the differential diagnosis remained broad. mNGS on CSF identified 839 unique read pairs (Table 3.2) that aligned to the genus *Cryptococcus*, essentially all of which aligned to *Cryptococcus neoformans*. Serum cryptococcal antigen was positive at a titer of > 1:160, and CSF cryptococcal antigen was positive at a titer of > 1:1280. Numerous fungal yeast forms were present in the CSF, and *Cryptococcus neoformans* grew in the CSF fungal culture. Except for hepatitis C virus, no other pathogens were identified via CSF mNGS (Figure 3.2) or by standard diagnostic assays. He clinically improved on additional anti-cryptococcal therapy.

#### *Recurrent Encephalopathy in the Context of Known HIV-1 Infection*

Participant 4 was a 55 year-old man with treatment-naïve HIV-1 infection diagnosed eight years prior. He presented to the Emergency Department with two weeks of confusion, decreased verbal output, as well as several months of cough, intermittent night sweats and weight loss; his CD4 count was 6 cells/mm<sup>3</sup>. Brain MRI showed bilateral, confluent and non-enhancing white matter T2 hyperintensities. CSF examination showed 0 WBC/mm<sup>3</sup>, 0 RBC/mm<sup>3</sup>, glucose 34 mg/dL, and total protein 93 mg/dL. CSF mNGS testing revealed more than 136,000 unique read pairs aligning to HIV-1 (Table 3.2, Figure 3.2). The full-length HIV-1 genome was assembled using paired read iterative contig extension (PRICE) v1.1.2.<sup>91</sup> The mean total read coverage across the HIV-1 genome was 3,991. No reads to JC virus, herpesviruses or fungal

pathogens were observed. His clinical course was consistent with HIV-1 dementia, and no other opportunistic CNS infection was identified with extensive testing.

### *Aspergillus oryzae*

Participant 5 was a 32 year-old man who presented to the Emergency Department with seven months of episodic dizziness, diplopia, headache, left facial numbness and weakness. He had a history of IDU and hepatitis C virus infection. HIV-1 testing was negative. Brain MRI revealed contrast enhancement in the left pons, middle cerebellar peduncle, and right posterior aspect of the pituitary infundibulum. A computed tomography (CT) angiogram revealed multiple areas of focal stenosis in the posterior circulation. CSF examination showed 95 WBC/mm<sup>3</sup> (77% lymphocytes, 13% neutrophils, 6% monocytes, 4% reactive lymphocytes), 0 RBC/mm<sup>3</sup>, total protein of 61 mg/dL, glucose of 45 mg/dL. Bacterial and fungal cultures were negative for CSF and blood. He was treated for a suspected autoimmune process with glucocorticoids. Two weeks later the patient's symptoms worsened. Repeat brain MRI revealed a new punctate infarct in the right thalamus. Repeat CSF examination showed worsening pleocytosis with a neutrophilic predominance. Bacterial cultures were negative. CSF *Aspergillus* galactomannan returned at 11.26 (positive > 0.5). A third CSF exam showed an even greater pleocytosis of 343 WBC/mm<sup>3</sup> (63% lymphocytes, 20% neutrophils, 15% monocytes, 2% reactive lymphocytes). CSF *Aspergillus* galactomannan was again elevated at 7.23. CSF mNGS returned 857 read-pairs mapping to *Aspergillus* spp. (Table 3.2) with the majority of the reads mapping to *Aspergillus oryzae*. CSF 18s rRNA PCR also returned positive for *Aspergillus* spp., and the CSF 1,3-β-D-glucan was



elevated. No cause of immunocompromise was identified, and no systemic aspergillus infection was identified. Treatment with oral voriconazole was initiated, and his hydrocephalus was treated with ventriculoperitoneal shunting. He was then lost to follow-up.

### *Histoplasma capsulatum*

Participant 6 was a 10 year-old girl with an early childhood spent in Indiana and Ohio, who presented in Seattle with six months of back pain and 10 days of progressive left facial droop, headache, neck pain, and vomiting. She had severe meningismus on exam. Brain and spinal MRI showed diffuse leptomeningeal enhancement. CSF examination revealed 68 WBC/mm<sup>3</sup> (33% neutrophils and >30% monocytes), 98 RBC/mm<sup>3</sup>, glucose <20 mg/dL, and total protein 292 mg/dL. Empiric therapy was initiated for bacterial meningitis, herpes simplex virus, and endemic mycoses. She developed respiratory failure and paraplegia. Repeat neuroimaging demonstrated new pontine infarcts and non-enhancing, longitudinal T2 hyperintensities throughout the cervicothoracic cord. Her therapy was broadened to cover TB meningitis. Ultimately, CSF fungal culture and 18s rRNA PCR revealed *Histoplasma capsulatum*. mNGS showed 33 read-pairs mapping to *Histoplasma capsulatum* (Table 3.2, Figure 3.2). After two months of treatment, she could ambulate but continues to have profound left hearing loss, facial weakness and neurogenic bladder.

### *Candida dubliniensis*

Participant 7 was a 26 year-old woman with an initially undisclosed history of IDU who presented with one year of atraumatic lower back pain followed by subacute

development of saddle anesthesia and left foot drop. On MRI, she had a loculated, rim-enhancing collection extending from the top of the lumbar spine anteriorly compressing the conus medullaris against the posterior wall, in addition to diffuse leptomeningitis involving the entire spinal cord and brainstem (Figure 3.3C-E). Cisternal CSF showed 126 WBC/mm<sup>3</sup> (67% neutrophils, 22% lymphocytes and 11% monocytes), 4 RBC/mm<sup>3</sup>, total protein of 105 mg/dL, and a glucose of 40 mg/dL. Extensive infectious disease diagnostic studies were unrevealing, and 11 weeks later the patient underwent lumbar meningeal biopsy. The pathology revealed non-inflammatory, dense fibrous tissue, and no microbes were identified. Consistent with the pathology that did not show evidence of active infection in sample that was biopsied, 18s rRNA and 16s rRNA PCRs, and mNGS of biopsy tissue did not reveal evidence of infection. 18s rRNA PCR of CSF was also negative. Three months later the patient became wheelchair-bound. Repeat cisternal CSF showed 700 WBC/mm<sup>3</sup> (81% neutrophils, 16% lymphocytes, 2% monocytes and 1% eosinophils), 2 RBC/mm<sup>3</sup>, total protein of 131 mg/dL and a glucose of 44 mg/dL. CSF mNGS revealed 68 read-pairs mapping to *Candida* spp. (Table 3.2) with 61 of the 68 pairs mapping to *Candida dubliniensis* with 99-100% identity. A CSF 1,3-β-D-glucan assay was 211 pg/mL (<80 pg/mL), whereas the serum 1,3-β-D-glucan assay had been repeatedly normal. Repeat CSF 18s rRNA and 16s rRNA PCRs were negative. The patient is being treated with combination anti-fungal therapy with mild clinical improvement, normalization of her CSF profile (including a negative CSF 1,3-β-D-glucan) and decreasing leptomeningeal enhancement on MRI. Two of the previous three reported cases of *Candida dubliniensis* meningitis were in patients with a history of IDU.<sup>92,93,94</sup>

### 3.3.2 Background Signature of Reagent and Environmental Contaminants

Examination of nucleotide alignments generated by non-templated water-only controls (n=24) and non-infectious CSF samples (n=94) revealed 4,400 unique bacterial, viral, and eukaryotic genera. This microbial background signature was predominated (>70%) by consistent proportions of bacterial taxa, primarily the Proteobacteria and Actinobacteria classes (Figure 3.4) representing common soil, skin, and environmental flora previously reported as laboratory and reagent contaminants.<sup>88</sup> To determine if these common microbial contaminants may have been misclassified as pathogens in previously published studies, we examined publicly available data from two cases of meningoencephalitis for which a possible infection was identified by mNGS.<sup>90</sup> In each case, neither organism (*Delftia acidovorans*, *Elizabethkingia*) was present at levels significantly greater than the mean of our background dataset of water-only and non-infectious CSF controls. We then examined data from a study aiming to characterize the “brain microbiome” and correlate brain dysbiosis to disease.<sup>89,10</sup> The abundance of the purported brain microbiota reveal distributions that are well within the observed variance we observe within our set of background water-only, non-templated controls (Figure 3.4). Of note, the authors of the brain microbiome study did not deep sequence water controls as they could not generate measurable quantities of DNA after reverse transcription-PCR (RT-PCR). The presence of environmental contaminants is due in part to low amounts of input RNA, which is frequently the case with acellular CSF samples, combined with the PCR amplification cycles necessary to generate a sequencing library. To assess this explicitly, we performed an RNA doping experiment (Figure 3.5) on a water sample and an uninfected CSF sample from which there was no

detectable cDNA after RT-PCR. The mNGS libraries made from the water and CSF samples had 9.4% and 7.6% unique, non-human sequences, respectively. The proportion of non-human sequences dropped dramatically after spiking in only 20 picograms of RNA of a known identity suggesting that non-human environmental sequences are particularly problematic for low input nucleic acid samples, which is often the case for CSF.

### 3.4 DISCUSSION AND CONCLUSIONS

We present seven diagnostically challenging cases of subacute and chronic meningitis in which mNGS of CSF identified a pathogen, including a case of subarachnoid neurocysticercosis that defied diagnosis for one year, the first case of CNS vasculitis caused by *Aspergillus oryzae*, and the fourth reported case of *Candida dubliniensis* meningitis. A straightforward statistical model leveraging a large mNGS dataset obtained from water-only, non-templated controls and patients with a variety of non-infectious neuroinflammatory syndromes correctly prioritized the pathogens. Larger, prospective studies are needed to determine the clinical utility of this approach for reducing the number of false positives and false negatives.

CSF mNGS has the potential to overcome several limitations of conventional CNS infectious disease diagnostics. First, the inherent risks of brain and/or meningeal biopsy make CSF mNGS a particularly attractive and less invasive diagnostic option for patients with suspected CNS infection. Second, the large number of neuroinvasive pathogens that cause subacute or chronic meningitis makes it logistically challenging and cost-prohibitive to order every possible neuro-infectious diagnostic test using a

candidate-based approach. Third, some assays lack sensitivity in the context of impaired immunity or acute infection (e.g., West Nile virus serology), can be slow to yield results (e.g., mycobacterial and fungal cultures) or may fail to differentiate between active infection and prior exposure (e.g., cysticercosis antibody or the interferon-gamma release assay test for *M. tuberculosis*).

The unbiased nature of mNGS makes the datasets inherently polymicrobial and complex. Thus, statistical scoring and filtering is essential to enhance the ability to discriminate between insignificant contaminants and true infectious organisms. Our algorithm correctly prioritized etiologic pathogens in these seven clinically-confirmed cases of infectious meningitis despite the fact that the pathogens ranged widely with regard to their absolute abundance (33-136,000 sequence read pairs) and the proportion of the non-human sequences (0.89-92.7%) that they comprised (Table 3.2).

In addition, we analyzed a recently published clinical mNGS dataset to highlight that a thoroughgoing profile of the microbes present in water-only controls and non-infectious CSF reinforces the skepticism with which the authors described a possible infection in one subject with *Delftia acidovorans* (patient 2) and in another subject with *Elizabethkingia* (patient 7) (Figure 3.4). Furthermore, such a database could help improve the accuracy of microbiome studies, especially for body sites historically considered sterile in which rigorous controls are necessary to establish that observed microbial sequences represent microbiota vs environmental contaminants (Supplementary Figure 1C).<sup>10,90</sup> This problem appears to be particularly acute in samples like CSF whose sub-nanogram levels of input RNA/DNA require unbiased molecular amplification steps before enough material is available for sequencing

applications. Indeed, the addition of only 20 picograms of purified RNA to a CSF sample was sufficient to suppress the majority of non-CSF reads deriving from the water and reagents (Figure 3.5). While amplifying the input signal increases the sensitivity of the assay, it also often over-represents the signature of contaminating taxa unique to a given laboratory, experimenter, or reagent lot.<sup>86,88</sup> These results provide a cautionary note and underscore the need for appropriate controls to aid in interpretation.

We expect larger databases of patient mNGS results will only enhance the ability to discriminate between irrelevant sequences and legitimate pathogens and permit more rigorous and probabilistic models for pathogen ranking and reporting. We present here one empirically derived system for prioritizing results, based on the read count weighted by standard z-scores. Given the sensitivity of NGS-based approaches, we anticipate individual laboratories will need to develop their own dynamic reference datasets to control for contaminants that are relevant to the particular time, place, and manner in which the biological samples are being analyzed.

mNGS represents an increasingly rapid and comparatively low cost means of screening CSF in an unbiased fashion for a broad range of human pathogens using a single diagnostic test. While this selected case series is not appropriate to measure the performance characteristics in a prospective cohort, a recently completed demonstration project sponsored by the State of California (<http://www.ciapm.org/project/precision-diagnosis-acute-infectious-diseases>), may also prove to be helpful in supporting the exclusion of CNS infection when a co-infection is suspected in an immunosuppressed patient (as illustrated by our cases with *C. neoformans* and HIV-1) or when a non-infectious cause, such as an autoimmune

condition, is clinically favored. On this basis, we foresee the eventual replacement of many single-agent assays performed in reference labs with a unified mNGS approach.

### **3.5 MATERIALS and METHODS**

Participants were recruited between September 2013 and March 2017 as part of a larger study applying mNGS to biological samples from patients with suspected neuroinflammatory disease. The seven participants had subacute or chronic leptomeningitis with or without encephalitis. An etiologic diagnosis was not known by the researchers at the time of study enrollment. If an infection was made by traditional means before mNGS testing was complete (participants 3, 5 and 6), the researchers performing mNGS remained blinded to the diagnosis. The cases were referred from the University of California, San Francisco (UCSF) Medical Center (n=2), Zuckerberg San Francisco General Hospital (n=2), Cleveland Clinic, University of Washington and Kaiser Permanente. The UCSF Institutional Review Board (IRB) approve the study protocol, and participants or their surrogates provided written informed consent. Treating physicians were informed about research-based mNGS results under an IRB-approved reporting mechanism.

#### **3.5.1 mNGS Protocol**

mNGS was performed on total RNA extracted from surplus CSF (250-500  $\mu$ L), and one subject also had mNGS performed on total RNA extracted from <50 mg of snap frozen, surplus tissue from a lumbar meningeal biopsy. Samples were processed for mNGS as previously described.<sup>77,79</sup> The non-human sequence reads have from

each sample have been deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive, BioProject (PRJNA338853).

### 3.5.2 Bioinformatics and Statistical Analysis

Paired-end 125-150 base pair (bp) sequences were analyzed using a previously described rapid computational pathogen detection pipeline consisting of open source components (Figure 1).<sup>77,79</sup> Unique, non-human sequences were assigned to microbial taxonomic identifiers (taxids) based on nucleotide (nt) and non-redundant (nr) protein alignments. To distinguish putative pathogens from contaminating microbial sequences derived from skin, collection tubes, lab reagents, or the environment, a composite background model of metagenomic data was employed. This model incorporated 24 water controls and 94 CSF samples from patients with non-infectious diagnoses, including 21 chronic meningitis with or without encephalitis cases. Data were normalized to unique reads mapped per million input reads (rpM) for each microbe at the species and genus level. Using this background dataset as the expected mean rpM for a given taxid, standard Z-scores were calculated for each genus (gs) and species (sp) in each sample based on the results from both the nt and nr database searches. Thus, there are four z-scores reported for each sample:  $Z_{nt,sp}$ ,  $Z_{nt,gs}$ ,  $Z_{nr,sp}$ ,  $Z_{nr,gs}$ . To prioritize reporting of the most unique (i.e., unexpected) taxa in each sample, the significance of each microbial species was mapped to a single value with the following empirically-derived formula:

$$\text{Score} = Z_{nt,sp} (Z_{nt,gs} (\text{rpM}_{nt})) + Z_{nr,sp} (Z_{nr,gs} (\text{rpM}_{nr}))$$



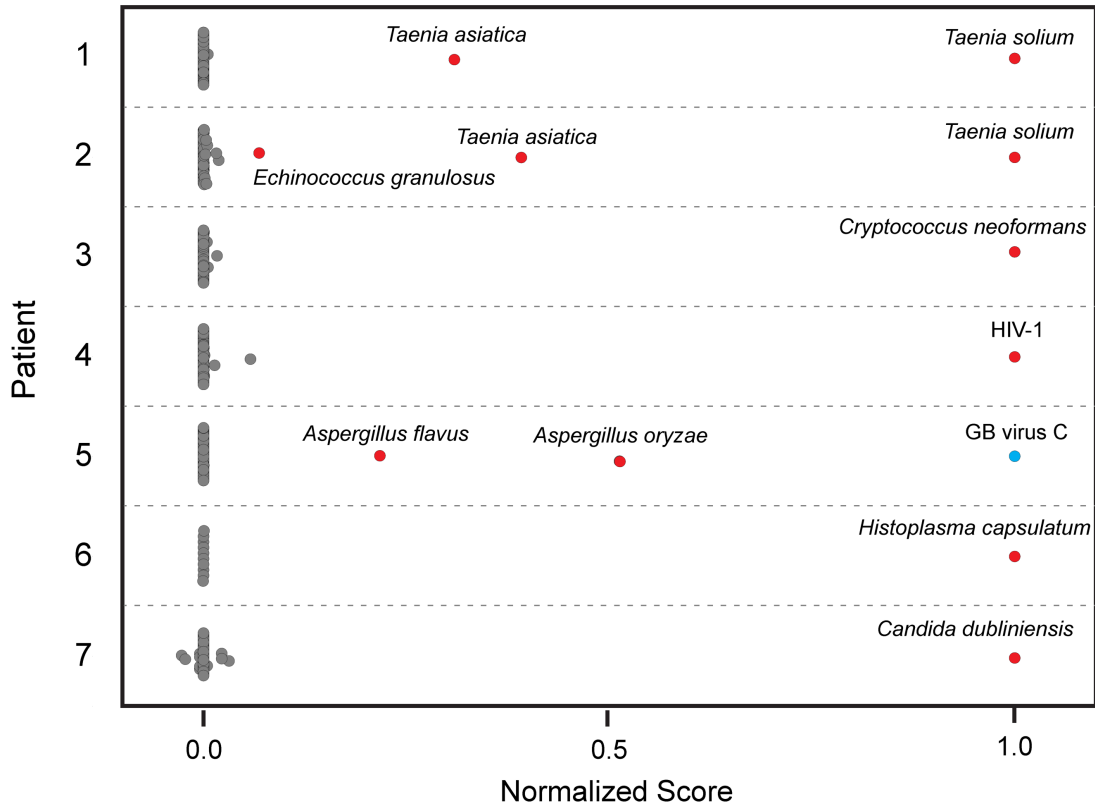
Here, the rpM values are scaled by both the z-score for the species and the genus. If both z-scores are negative, the product remains negative. The maximum z-score is arbitrarily capped at 100. This product is calculated for alignments to both the nt and nr databases and summed. The top-ranked taxa were considered with respect to the clinical context of the patient. Microbes with known CNS pathogenicity that could cause a clinical phenotype concordant with the clinical presentation were considered potential pathogens and were confirmed by standard microbiologic assays, as described in the brief case histories herein.

### 3.6 FIGURES

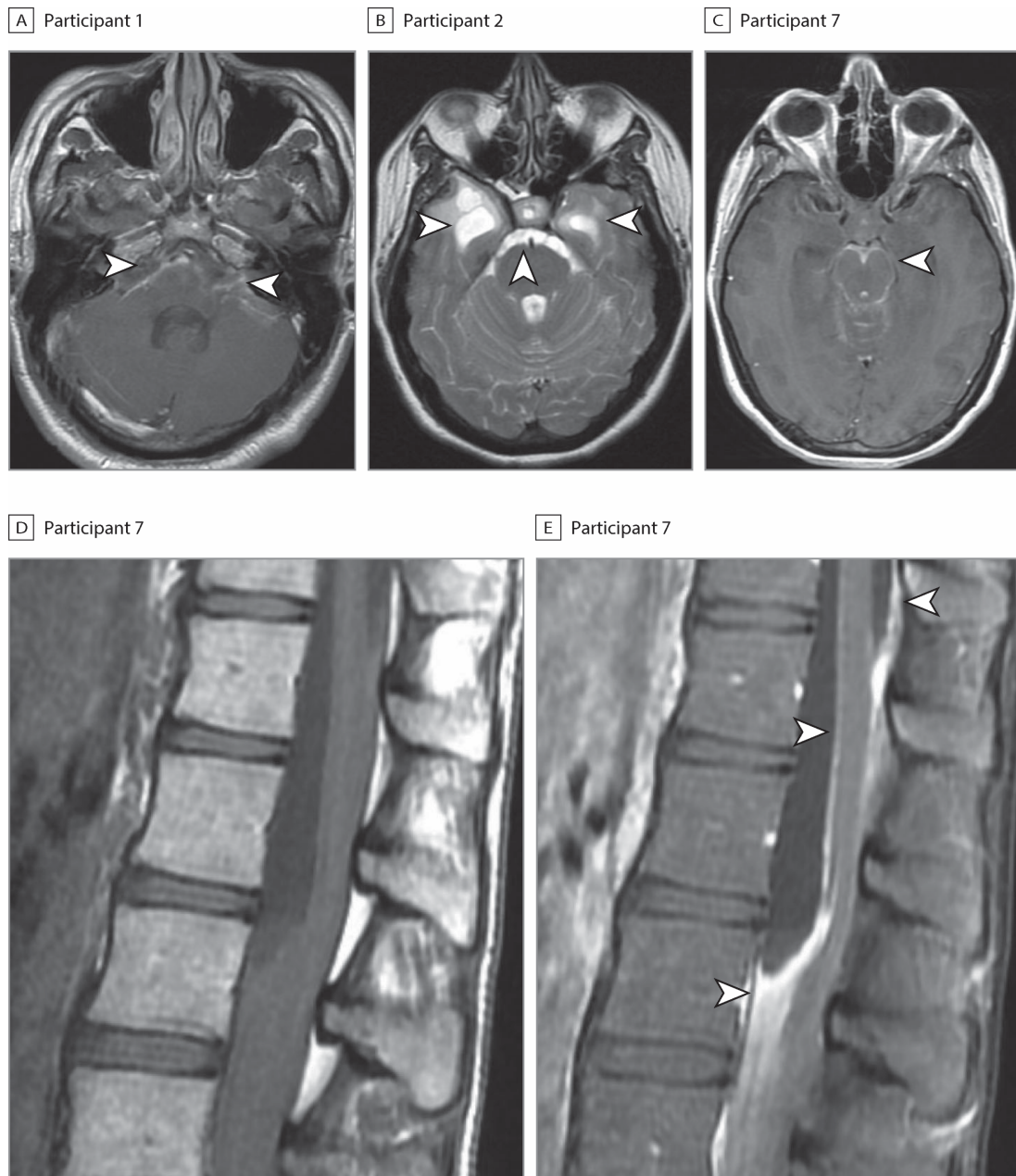
	Read-pairs, No.	Retained, %	Time, min	Component	Target
Raw sequence, .fastq files	13 141 550	100	0.0	NA	NA
First pass removal of human reads	1 930 493	14.7	7.2	STAR	Hg38/PanTro4 RepBase
Quality filter	538 661	4.1	7.9	PriceSeqFilter	Read-pairs
Compression of redundant reads	374 592	2.8	8.1	CD-HIT-DUP	Read-pairs
LZW complexity filter	347 719	2.6	9.5	LZW (script)	Read-pairs
Second pass paired-end human read removal	63 435	0.5	10.9	Bowtie 2	Hg38/RepBase
Alignment to nt database	62 855	0.5	12.3	GMAP/GSNAP	NCBI nt
Alignment to nr database	62 514	0.5	19.5	RAPSearch2	NCBI nr
Taxonomic statistics/reporting	58 789	94.0 of Nonhost	19.6	PHP/MySQL	NCBI taxonomy

↘ Aligns to GenusTaenia

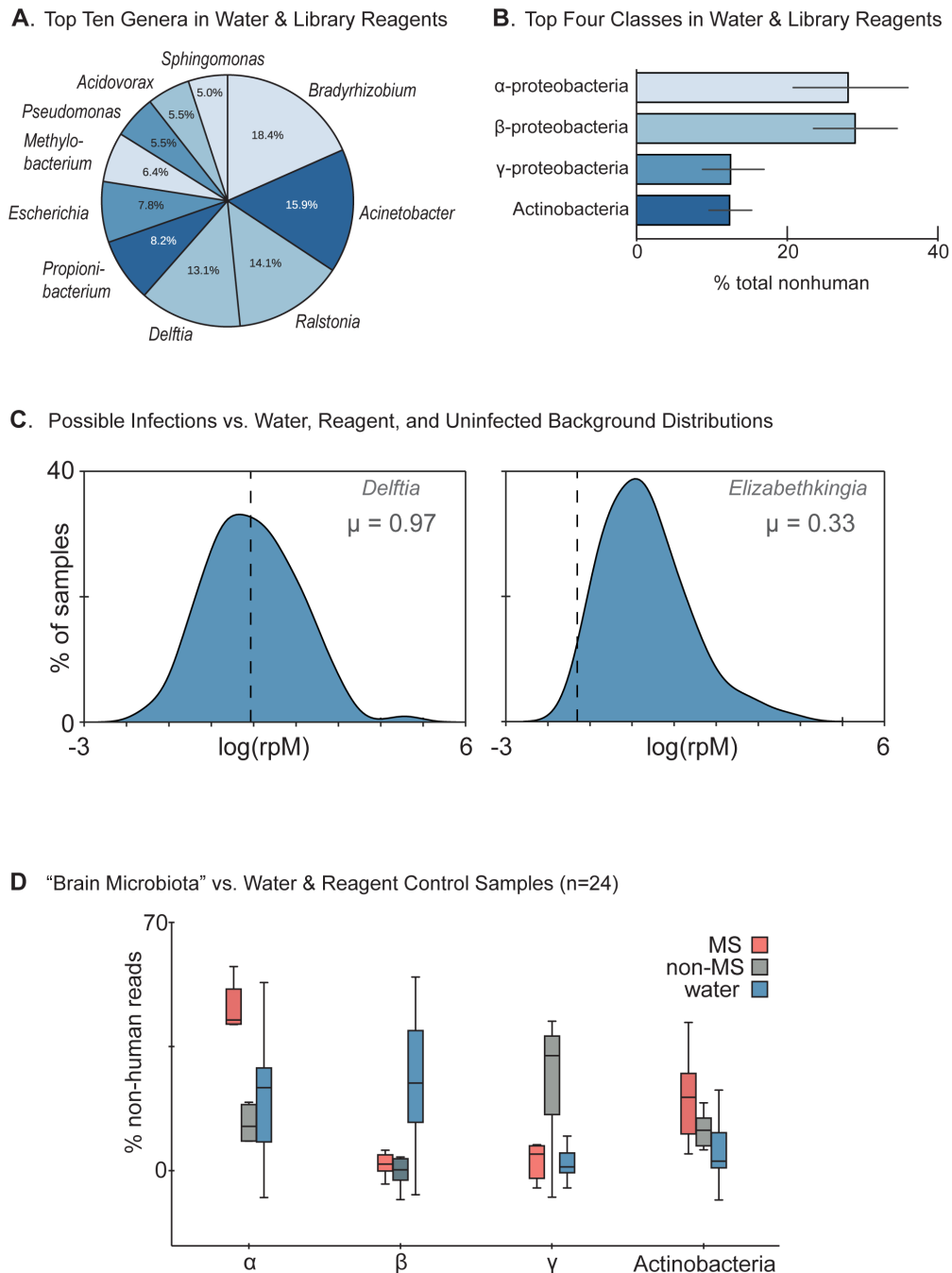
**Figure 3.1 - Diagram of a rapid computational pipeline.** Values for participant 1 are shown as an example. The cerebrospinal fluid sample obtained from participant 1 yielded 13,141,550 million read-pairs, which were then subjected to removal of human reads by spliced transcripts alignment to a reference (STAR, version 2.4.2),<sup>95</sup> quality control filtering (PriceSeqFilter, version 1.1.2),<sup>91</sup> compression of duplicate reads (CD-HIT-DUP, version 4.6.4-2015),<sup>96</sup> removal of low-complexity sequences by filtering for high Lempel-Ziv-Welch (LZW)<sup>81</sup> compression ratios, a second round of removal of human reads (Bowtie 2, version 2.2.4),<sup>60</sup> alignment to the National Center for Biotechnology Information (NCBI) nucleotide (nt) database (GMAP/GSNAP, version 2015-12),<sup>98</sup> alignment to the NCBI nonredundant (nr) protein database (RAPSearch2, version 2.23),<sup>99</sup> and statistical calculation and taxonomy reporting using PHP/MySQL, version 5.5.53. The entire computational pipeline was completed in 19.6 min using a single high-end server (32 core, Intel Xeon E5-2667 v3 with a 3.2-GHz processor and 768 Gb of RAM). NA indicates not applicable.



**Figure 3.2 - Ranked Results of Statistical Scoring.** Strip plot of normalized species significance scores for microbial taxa (colored circles) in each participant sample (row). In 6 of 7 samples, the neurologic infection (orange circles) is ranked as the most significant by our approach. In participant number 5, *Aspergillus oryzae* is ranked second behind GB virus C, a likely concurrent infection unassociated with the clinical presentation (blue circle). Microbes likely representing environmental contaminants are also shown (gray circles).

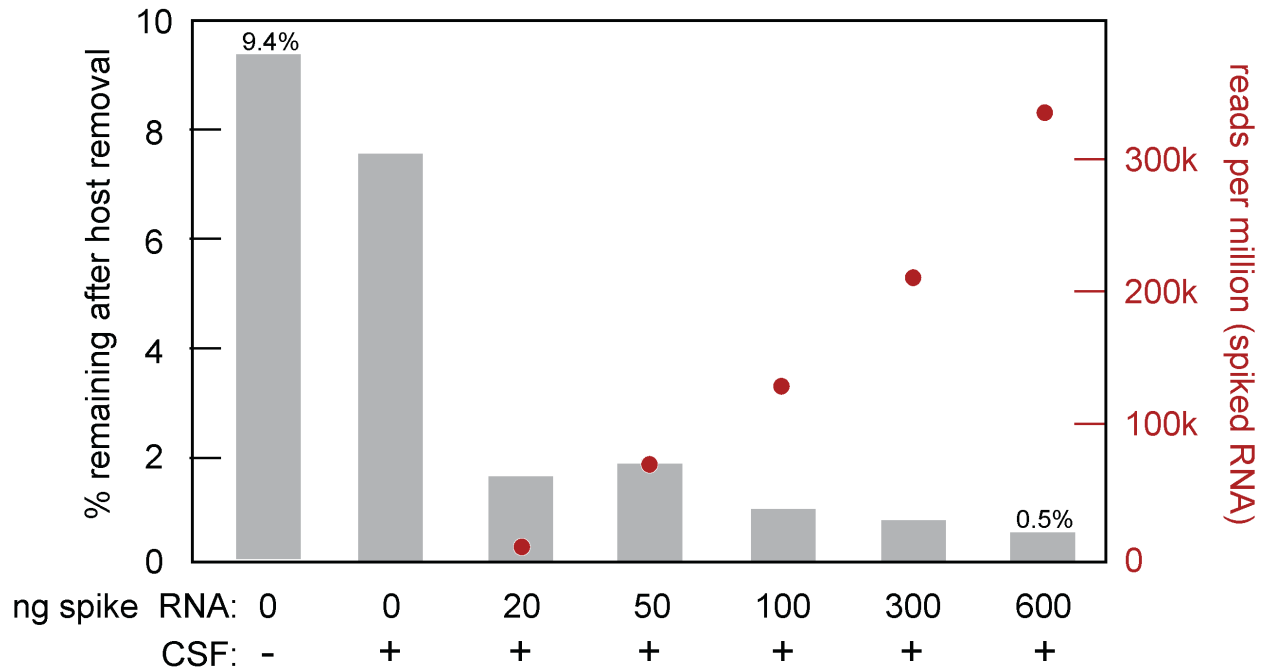


**Figure 3.3 - Selected Neuroimaging.** A.) Axial T1-weighted brain magnetic resonance image (MRI) with contrast enhancement demonstrating basilar meningitis (arrowheads) in a 28-year-old man (participant 1) with neurocysticercosis identified by metagenomic next-generation sequencing (mNGS). B.) Axial T2-weighted brain MRI demonstrating right anterior temporal lobe and prepontine cysts (arrowheads) in a 34-year-old woman with neurocysticercosis (participant 2) identified by mNGS. C-E.) Axial T1-weighted MRI with contrast enhancement showing basilar meningitis (C [arrowhead]), and a sagittal T1-weighted lumbar spine MRI showing a loculated rim-enhancing collection extending from the top of the lumbar spinal cord anteriorly and compressing the conus medullaris against the posterior wall without (D) and with (E, arrowheads) contrast in a 26-year-old woman with *Candida dubliniensis* meningitis (participant 7) identified by mNGS.



**Figure 3.4 - Background signature of reagent and environmental contaminants.** A.) The 10 most abundant genera identified water-only and reagent controls (n=24). 10 bacterial taxa account for ~50% of all non-human (viral, bacterial, fungal, and selected eukaryotes) sequences in the control data. B.) Organized at the class-level, four classes of bacteria ( $\alpha$ -proteobacteria,  $\beta$ -proteobacteria,  $\gamma$ -proteobacteria, Actinobacteria) represent >80% of the sequences in the wateronly and reagent controls. C.) The proportion of *Delftia acidovorans* sequences (dotted line at 0.93) CNS infection identified by mNGS is concordant with the expected mean of our control dataset (0.97). Similarly, the proportion of *Elizabethkingia* sequences (dotted line) in another

possible infectious case from the same report is significantly lower than the mean abundance in our control dataset. D.) Comparison of the observed variance of the four classes of bacteria (B) to publicly available data from a recent report<sup>10</sup> on microbiota in healthy and multiple sclerosis patient brain specimens suggests that the relative abundances of each taxa in the brain specimens are within the observed variance of our background dataset.



**Figure 3.5 - RNA doping experiment.** Comparison of the percent non-human sequences (y-axis) found in water (column 1) and a cerebrospinal fluid (CSF) control (column 2) and the decrease in the percent non-human sequences found with increasing amounts of spiked RNA of a known identity (columns 3-7). Added RNA was generated by T7 in vitro transcription from a cloned luciferase reporter gene, purified, quantified, and spiked into the CSF at the indicated amounts. These data suggest that common environmental contaminants are present at low picogram quantities, and the addition of only 20 pg is sufficient to suppress the majority of reads not derived from the CSF.

### 3.7 TABLES

Participant No./Sex/Age, y	Final Diagnosis	Clinical Presentation	Disease Duration at Time of Diagnostic LP, mo	Length of Follow-up, mo
1/M/28	Taenia solium	Headache, diplopia	11	8
2/F/34	T solium	Headache, unilateral facial numbness and tinnitus, recurrent loss of consciousness	9	21
3/M/52	Cryptococcus neoformans	Seizures, coma	3	19
4/M/55	HIV-1	Dementia, cough, night sweats, weight loss	0.5	7
5/M/32	Aspergillus oryzae	Headache, dizziness, diplopia, unilateral facial weakness and numbness	7	1
6/F/10	Histoplasma capsulatum	Back pain followed by unilateral facial droop, headache, neck stiffness	6	3.5
7/F/26	Candida dubliniensis	Low back pain followed by saddle anesthesia and unilateral foot drop	20	1

**Table 3.1 - Clinical characteristics of study participants**

Participant No.	Pathogen	Paired-End Read-Pairs, Total No.	Unique Nonhuman Nonredundant Read-Pairs, No.	Unique Pathogen Read-Pairs, No. (% Nonhuman Read-Pairs)
1	Taenia solium	13 141 550	63 435	58 789 (92.7)
2	T solium	17 712 171	3732	569 (15.2)
3	Cryptococcus neoformans	11 121 312	1678	839 (50.0)
4	HIV-1	8 529 421	261 847	136 000 (51.9)
5	Aspergillus oryzae	14 698 597	2753	857 (31.1)
6	Histoplasma capsulatum	13 385 787	9406	33 (0.4)
7	Candida dubliniensis	14 726 483	7636	68 (0.9)

**Table 3.2 - Metagenomic sequencing summary**



## Chapter 4

# Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications

### 4.1 ABSTRACT

Next-generation sequencing has generated a need for a broadly applicable method to remove unwanted high-abundance species prior to sequencing. We introduce DASH (Depletion of Abundant Sequences by Hybridization). Sequencing libraries are 'DASHed' with recombinant Cas9 protein complexed with a library of guide RNAs targeting unwanted species for cleavage, thus preventing them from consuming sequencing space. We demonstrate a more than 99% reduction of mitochondrial rRNA in HeLa cells, and enrichment of pathogen sequences in patient samples. We also demonstrate an application of DASH in cancer. This simple method can be adapted for any sample type and increases sequencing yield without additional cost.

## 4.2 INTRODUCTION

The challenge of extracting faint signals from abundant noise in molecular diagnostics is a recurring theme across a broad range of applications. In the case of RNA sequencing (RNA-Seq) experiments specifically, there may be several orders of magnitude difference between the most abundant species and the least. This is especially true for metagenomic analyses of clinical samples like cerebrospinal fluid (CSF), whose source material is inherently limited,<sup>75</sup> making enrichment or depletion strategies impractical or impossible to employ prior to library construction. The presence of unwanted high-abundance species, such as transcripts for the 12S and 16S mitochondrial ribosomal RNAs (rRNAs), effectively increases the cost and decreases the sensitivity of counting-based methodologies.

The same issue affects other molecular clinical diagnostics. In cancer profiling, the fraction of the mutant tumor-derived species may be vastly outnumbered by wild-type species due to the abundance of immune cells or the interspersed nature of some tumors throughout normal tissue. This problem is profoundly exaggerated in the case of cell-free DNA/RNA diagnostics, whether from malignant,<sup>100,101</sup> transplant,<sup>102</sup> or fetal sources,<sup>103,104</sup> and relies on brute force counting by either sequencing or digital PCR (dPCR)<sup>105</sup> to yield a detectable signal. For these applications, a technique to deplete specific unwanted sequences that is independent of sample preparation protocols and agnostic to measurement technology is highly desired

CRISPR (clustered regularly interspaced short palindromic repeats) and Cas (CRISPR associated) nucleases, such as Cas9, function in bacterial adaptive immune systems to remove infecting phage DNA from the host without harm to the bacteria's

own genome. The CRISPR-Cas9 system has attained widespread adoption as a genome editing technique.<sup>106,107,108,109</sup> When coupled with single guide RNAs (sgRNAs) designed against targets of interest, *Streptococcus pyogenes* Cas9 binds to 3' NGG protospacer adjacent motif (PAM) sites and produces double-stranded breaks if the sgRNA successfully hybridizes with the adjacent target sequence (Figure 4.1). In vitro, Cas9 may be used to cut DNA directly, in a manner analogous to a conventional restriction enzyme, except that the target sequence (outside of the PAM site) may be programmed at will and massively multiplexed without significant off-target effects. This affords the unique opportunity to target and prevent amplification of undesired sequences, such as those that are generated during next-generation sequencing (NGS) protocols.

In this chapter, we have exploited the unique properties of Cas9 to selectively deplete unwanted high-abundance sequences from existing RNA-Seq libraries. We refer to this approach as Depletion of Abundant Sequences by Hybridization (DASH). Employing DASH after transposon-mediated fragmentation (or adapter ligation) but prior to the following amplification step (which relies on the presence of adaptor sequences on both ends of the fragment) prevents amplification of the targeted sequences, thus ensuring they are not represented in the final sequencing library (Figure 4.1b). We show that this technique preserves the representational integrity of the non-targeted sequences while increasing overall sensitivity in cell line samples and human metagenomic patient samples. Further, we demonstrate the utility of this system in the context of cancer detection, in which depletion of wild-type sequences increases the detection limit for oncogenic mutant sequences. The DASH technique may be used to

deplete specific unwanted sequences from existing Illumina sequencing libraries, PCR amplicon libraries, plasmid collections, and virtually any other existing collection of DNA species.

Existing specific sequence enrichment techniques — such as pull-down methods,<sup>103,110,111,112</sup> amplicon-based methods,<sup>101,113</sup> molecular inversion methods,<sup>114,115,116</sup> COLD-PCR,<sup>117</sup> competitive allele-specific TaqMan PCR (castPCR),<sup>118</sup> and the classic method of using restriction enzyme digestion on mutant sites — can effectively enrich for targets in sequencing libraries,<sup>119</sup> but these are not useful for discovery of unknown or unpredicted sequences. Brute force counting methods also exist, such as dPCR,<sup>101,105</sup> but they are not easy to multiplex across a large panel of samples. While high-throughput sequencing of select regions can be highly multiplexed to detect rare and novel mutations, and barcoded unique identifiers can overcome sequencing error noise,<sup>120</sup> it is costly since the vast majority of the sequencing reads map to non-informative wild-type sequences. A number of sequence-specific RNA depletion methods also currently exist. Illumina's Ribo-Zero rRNA Removal Kit and Ambion's GLOBINclear Kit pull rRNAs and globin mRNAs, respectively, out of total RNA samples using sequence-specific oligos conjugated to magnetic beads. RNase H-based methods, such as New England BioLab's NEBNext rRNA Depletion Kit similarly mark abundant RNA species with sequence-specific DNA oligos, and then subject them to degradation by RNase H, which digests RNA/DNA hybrid molecules.<sup>121</sup> These methods are all employed prior to the start of library prep, and are limited to samples containing at least 10 ng to 1 µg of RNA. DASH, in contrast, depletes abundant species after

complementary DNA (cDNA) amplification, and thus can be utilized for essentially any amount of input sample.

### 4.3 RESULTS

We demonstrate deletion of unwanted mitochondrial rRNA using DASH first on HeLa cell line RNA (Figure 4.2) and then on CSF RNA from patients with pathogens in their CSF (Figure 4.3), in order to increase sequencing bandwidth of useful data. Selection of rRNA sgRNA targets was based on examining coverage plots for standard RNA-Seq experiments on HeLa cells as well as on several patient CSF samples. Coverage of the 12S and 16S mitochondrial rRNA genes was consistently several orders of magnitude higher than the rest of the mitochondrial and non-mitochondrial genes (Figs. 4.2c and 4.3). We chose 54 sgRNA target sites within this high-coverage region of the mitochondrial chromosome, situated approximately every 50 bp over a 2.5 kb region (Table 4.2). sgRNA sites are indicated by red arrowheads in Fig. 4.2B. sgRNAs for these sites were generated as described in the "Materials and methods" section.

To calculate the input ratio of Cas9 and sgRNA to sample nucleic acid, we estimated (based on prior experience generating mNGS libraries from human samples) that 90% of each sample was comprised of the rRNA regions that we targeted; thus, our potential substrate makes up 4.5 ng of a 5 ng sample. This corresponds to a target site concentration of 13.8 nM in the 10  $\mu$ L reaction volume. To assure the most thorough Cas9 activity possible, and given that Cas9 is a single-turnover enzyme in vitro,<sup>122</sup> we used a 100-fold excess of Cas9 protein and a 1000-fold excess of sgRNA relative to the

target. Thus, each 10  $\mu$ L sample of cDNA generated from a CSF sample contained a final concentration of 1.38  $\mu$ M Cas9 protein and 13.8  $\mu$ M sgRNA. In the case of HeLa cDNA, we used only 1 ng per sample, and therefore decreased the Cas9 and sgRNA concentrations by fivefold. However, since mitochondrial rRNA sequences represented only approximately 60% of the HeLa samples (compared with approximately 90% for CSF), the HeLa samples contained 150-fold Cas9 and 1500-fold sgRNA. To examine dose response, we processed additional 1 ng HeLa samples treated with 15-fold Cas9 and 150-fold sgRNA. Both concentrations were done in triplicate (Figure 4.1).

#### **4.3.1 Reduction of unwanted abundant sequences in HeLa samples**

We first demonstrate the utility and efficacy of our approach using sequencing libraries prepared from total RNA extracted from HeLa cells. In the untreated samples, reads mapping to 12S and 16S mitochondrial rRNA genes represent 61% of all uniquely mapped human reads. After DASH treatment, these sequences are reduced to only 0.055% of those reads (Figure 4.2). Comparison of gene-specific fragments per kilobase of transcript per million mapped reads (fpkm) values between treated and untreated samples reveals mean 82-fold and 105-fold decreases in fpkm values for 12S and 16S rRNA, respectively, in the samples treated with 150-fold Cas9 and 1500-fold sgRNA. Similarly, the samples treated with 15-fold Cas9 and 150-fold sgRNA show 30 and 45-fold reductions in 12S and 16S fpkm values, respectively, indicating a dose-dependent response to DASH treatment (Figure 4.1).

### 4.3.2 Enrichment of non-targeted sequences in HeLa samples

This profound depletion of abundant 12S and 16S transcripts increases the available sequencing capacity for the remaining, untargeted transcripts. We quantify this increase by the slope of the regression line fit to the remaining genes, showing a 2.38-fold enrichment in fpkm values for all untreated transcripts. An  $R^2$  coefficient of 0.979 for this regression line indicates strong consistency between replicates with minimal off-target effects (Figure 4.2c).

To confirm that our depletion was specific to only the targeted mitochondrial sequences, we calculated the changes in fpkm values across all genes in the treated and untreated samples and identified those genes that were significantly diminished ( $>2$  standard deviations) relative to their control values. To overcome issues with stochastic variation at low gene counts/fpkm, we eliminated those genes that, between the three technical replicates at each Cas9 concentration, showed standard deviations in fpkm values greater than 50% of the mean. All of the genes meeting this criterion were present at less than 15 fpkm. Of the remaining genes, only one non-targeted human gene, MT-RNR2-L12, showed significant depletion when compared with the un-treated samples (Fig. 4.2c). MT-RNR2-L12 is a pseudogene and shares over 90% sequence identity with a portion of the 16S mitochondrial rRNA gene. Out of the 24 sgRNA sites within the homologous region, 16 of them retain intact PAM sites in MT-RNR2-L12. Of these, seven have perfectly matching 20mer sgRNA target sites, and the remaining nine each have between one and four mutations (Figure 4.6). Depletion of this gene is, therefore, an expected consequence of our sgRNA choices.

### 4.3.3 Reduction of unwanted abundant sequences in CSF samples

We next tested the utility of our method when applied to clinically relevant samples. In the case of pathogen detection in patient samples, the microbial transcripts are typically low in number and become greatly outnumbered by human host sequences. As a result, sequencing depth must be drastically increased to confidently detect such small minority sequence populations. We reasoned that depletion of unwanted high-abundance sequences from patient libraries could result in increased representation of pathogen-specific sequence reads. We thus integrated the DASH method with our in-house metagenomic deep sequencing diagnostic pipeline for patients with meningeal inflammation (i.e., meningitis) or brain inflammation (i.e., encephalitis) likely due to an infectious agent or pathogen. Figure 4.3 and Table 4.1 summarize the results of this analysis. In all three cases, the DASHed and untreated samples have a similar number of reads (1.8–3.4 million), but DASHing reduces the number of duplicate reads, indicating an increase in library complexity.

In the case of a patient with meningoencephalitis whose CSF was previously shown to be infected with the amoeba *Balamuthia mandrillaris* (patient 1),<sup>76</sup> diagnosis was originally made by identification of a small fraction (<0.1%) of reads aligning to specific regions of the *B. mandrillaris* 16S mitochondrial gene. After DASH treatment, human mitochondrial 12S and 16S genes were reduced by more than an order of magnitude, and sequencing coverage of the *B. mandrillaris* 16S fragment increased 3.6-fold. Notably, *B. mandrillaris* is a eukaryotic organism, yet depletion of the human 16S gene by DASH did not have off-target effects on the 16S *B. mandrillaris* mitochondrial gene. Similarly, patient CSF samples with confirmed *Cryptococcus neoformans* (fungus;



patient 2) and *Taenia solium* (pork tapeworm; patient 3) infections showed 2- and 3.9-fold increases in coverage of the 18S genes of *C. neoformans* and *T. solium*, respectively, the detection of which was crucial in the initial diagnoses. The observed increases in relative signal can be translated into either a sequencing cost savings or a higher sensitivity that may be useful clinically for earlier detection of infections.

#### **4.3.4 Reduction of wild-type background for detection of the KRAS G12D (c.35G>A) mutation in human cancer samples**

Specific driver mutations known to promote cancer evolution and at times to make up the genetic definition of malignant subtypes are important for diagnosis and targeted therapeutics. In complex samples isolated from biopsies or cell-free body fluids such as plasma, wild-type DNA sequences often overwhelm the signal from mutant DNA, making the application of traditional Sanger sequencing challenging.<sup>100,101,123</sup> For NGS, detection of minority alleles requires additional sequencing depth and increases cost. We reasoned that the DASH technique could be applied to increase mutation detection from a PCR amplicon derived from a patient sample. We chose to focus on depletion of the wild-type allele of KRAS at the glycine 12 position, a hotspot of frequent driver mutations across a variety of malignancies.<sup>124,125,126</sup> This is an ideal site for DASH because all codons encoding the wild-type glycine residue contain a PAM site (NGG), while any mutation that alters that residue (e.g., c.35G>A, p.G12D) ablates the PAM site and is thus uncleavable by Cas9 (Figure. 4.4a). This will be true of any mutation that changes a glycine (codons GGA, GGC, GGG, and GGT) or a proline (codons CCA, CCC, CCG, and CCT) to any other amino acid. Furthermore, it is relevant to the ubiquitous C>T nucleotide change found in germline mutations as well as somatic

cancer mutations.<sup>127</sup> Targeting of other mutations will likely be possible in the near future with reengineered CRISPR nucleases or those that come from alternative species and have different PAM site specificities.<sup>128,129</sup>

The sequence of the sgRNA designed to target the KRAS G12D PAM site is listed in Table 4.2, as is the non-human sequence used for the negative control sgRNA. Both were transcribed from a DNA template by T7 RNA polymerase, purified, and complexed with Cas9 as described in the "Materials and methods" section. Samples were prepared by mixing sheared genomic DNA from a healthy individual (with wild-type KRAS genotype confirmed with dPCR) and KRAS G12D genomic DNA to achieve mutant to wild-type allelic ratios of 1:10, 1:100, and 1:1000, and 0:1. For each mixture, 25 ng of a DNA was incubated with 25 nM Cas9 pre-complexed with 25 nM of sgRNA targeting KRAS G12D. This concentration is high relative to the concentration of target molecules, but empirically we found it to be the most efficient ratio. We hypothesize that this may be due to non-cleaving Cas9 interactions with the rest of the human genome,<sup>122</sup> which effectively reduce the Cas9 concentration at the cleavage site.

Samples were subsequently heated to 95 °C for 15 min in a thermocycler to deactivate Cas9 ("Materials and methods"). Droplet digital PCR (ddPCR) was used to count wild-type and mutant alleles using the primers and TaqMan probes depicted in Fig. 4.4a and described in the "Materials and methods" section. All samples were processed in triplicate. Samples incubated with or without Cas9 complexed to a non-human sgRNA target show the expected percentages of mutant allele: approximately 10 %, 1 %, and 0.1 % for the 1:10, 1:100, and 1:1000 initial mixtures respectively (Fig. 4.4b). With addition of Cas9 targeted to KRAS, the wild-type allele count drops nearly

two orders of magnitude (purple bars in Fig. 4.4b), while virtually no change is observed in number of mutant alleles (blue bars). This confirms the high specificity of Cas9 for the NGG of the PAM site.

With the addition of DASH targeted to KRAS G12, the percentage of mutant allele jumps from 10 % to 81 %, from 1 % to 30 %, and from 0.1 % to 6 % (Fig. 4.4c). This corresponds to 8.1-fold, 30-fold and 60-fold representational increases for the mutant allele, respectively. As expected, there was virtually no detection of mutant alleles in the wild-type-only samples both with and without DASH treatment (one droplet in one of three no DASH wild-type-only samples).

#### **4.4 DISCUSSION**

In this chapter I have described DASH, a technique that leverages in vitro Cas9 ribonucleoprotein (RNP) activity to deplete specific unwanted high-abundance nucleotide sequences, which results in the enrichment of rare and less abundant sequences in NGS libraries or amplicon pools.

While the procedure may be easily generalized, we developed DASH to address current limitations in metagenomic pathogen detection and discovery, where the sequence abundance of an etiologic agent may be present as a minuscule fraction of the total. For example, infectious encephalitis is a syndrome caused by well over 100 pathogens ranging from viruses, fungi, bacteria and parasites. Because of the sheer number of diagnostic possibilities and the typically low pathogen load present in CSF, more than half of encephalitis patients never have an etiologic agent identified.<sup>130</sup> We have demonstrated that NGS is a powerful tool for identifying infections, but as the *B.*

*mandrillaris* meningoencephalitis case demonstrates, the vast majority of sequence reads are “wasted” re-sequencing high abundance human transcripts. In this case, we have shown that DASH depletes with incredible specificity the small number of human rRNA transcripts that comprise the bulk of the NGS library, thereby lowering the required sequencing depth to detect non-human sequences and enriching the proportion of non-human (*Balamuthia*) reads in the metagenomic dataset. In this study, we have targeted mitochondrial rRNA species because we have consistently observed them to be the most abundant sequences in these CSF-derived RNA samples. For other types of tissues, alternative programming of DASH for removal of nuclear rRNA species or essentially any other abundant sequences would be warranted.

In the case of infectious agents, it is possible to directly enrich rare sequences by hybridization to DNA microarrays or beads.<sup>110,131</sup> However, these approaches rely on sequence similarity between the target and the probe and therefore may miss highly divergent or unanticipated species. Furthermore, the complexity and cost of these approaches will continue to increase with the known spectrum of possible agents or targets. In contrast, the identity and abundance of unwanted sequences in most human tissues and sample types has been well described in scores of previous transcriptome profiling projects,<sup>121</sup> and therefore optimized collections of sgRNAs for DASH depletion are likely to remain stable.

A number of methods for depleting ribosomal RNA from RNA-Seq libraries exist in the form of commercially available kits. We assert that DASH is equally effective or better than these methods on four metrics: (1) input requirements, (2) performance, (3) programmability, and (4) cost. These can be assessed based on information available

on company websites or in publications for three major competing techniques: Illumina's Ribo-Zero and Thermo Fisher's RiboMinus, which both use biotinylated capture probes for depletion; and New England Biolab's NEBNext rRNA depletion kit, which uses RNase H for depletion.

#### **4.4.1 Input requirements**

Illumina recommends 1 µg of total RNA as input for Ribo-Zero, but also has a low-input protocol requiring only 100 ng. ThermoFisher recommends 2–10 µg of total RNA for its standard RiboMinus protocol, and 100 ng to 1 µg for its Low Input RiboMinus Eukaryote System v.2. NEB recommends 10 ng to 1 µg total RNA input for the NEBNext rRNA Depletion Kit. The reason for these stringent amount requirements is that these three methods all deplete samples at the RNA stage. DASH, in contrast, avoids the need to delicately manipulate the original sample. Instead, DASH is employed after cDNA synthesis and library generation; thus, it can be performed on any library, without regards to starting total RNA amount, or the manner in which the library was constructed (tagmentation or otherwise). For scarce and precious samples, such as patient CSF, often less than 10 ng of total cDNA is available even after NuGEN Ovation amplification; prior to this work, no commercial depletion method was available for these samples.

#### **4.4.2 Performance**

All commercial rRNA depletion methods promise at least 85% reduction in reads of the sequences they target. Illumina states that the Ribo-Zero technique can achieve

between 85% and >99% reduction in the rRNA sequences it targets; RiboMinus states 95–98 % reduction; and NEBNext states 95–99% reduction. Adiconis et al. compared several RNA-Seq methods and reported on many metrics, including depletion of rRNA sequences.<sup>121</sup> Ribosomal RNA sequences comprised 84.7% of reads in their undepleted sample (100 ng total RNA from K-562 cells), while Ribo-Zero reduced this to 11.3% (an 86.7% reduction), and RNase H reduced it to 0.1% (a 99.9% reduction). In this paper, we show that DASH decreases the mitochondrial rRNA reads in HeLa total RNA from 61% to 0.055% (99.9% reduction). Adiconis et al. obtained similar numbers from 1 µg total RNA samples from formalin-fixed paraffin-embedded (FFPE) kidney tissue (78.2 % and 99.9 % reduction for Ribo-Zero and RNase H, respectively) and pancreas tissue (73.0 % and 99.7 % reduction for Ribo-Zero and RNase H, respectively). This is comparable to DASH reduction in three patient CSF samples (82.1%, 81.4% and 88.2% reduction). However, it is important to note again that Adiconis et al. used 1 µg total RNA from tissue samples, while the DASHed CSF samples consisted of only 5 ng of NuGEN Ovation-amplified cDNA (total RNA content in the original CSF samples was too low to accurately quantify).

Another important measure of performance is maintenance of relative abundances of non-targeted sequences, such as the human transcriptome. Correlation coefficients for samples with and without DASH treatment ranged from  $R^2 = 0.979$  to 0.994 in this study (Figures 4.2, Figure 4.6), slightly higher than those found by Adiconis et al. for all methods.

### **4.4.3 Programmability**

DASH can be adapted to target any sequence containing a PAM site; construction of new sgRNAs is facile and inexpensive (see "Materials and methods" section). Because it is employed after sequencing adapter addition, DASH's utility is not limited to RNA-Seq; it can be applied to any library type. Examples include ATAC-Seq libraries, in which desired nuclear DNA is contaminated with a significant amount of mitochondrial DNA sequences, and microbiome sequencing, where it may be desirable to eliminate a particularly abundant species in order to better sample the underlying diversity. Since Ribo-Zero, RiboMinus and NEBNext are all proprietary kits, they cannot easily be re-programmed by the user to target other sites.

### **4.4.4 Cost**

Based on current publicly available list prices of the most economical kit sizes, the per-sample costs (in US dollars) of the kits discussed here are \$82.00 (Ribo-Zero Gold Kit H/M/R), \$93.67 (RiboMinus Human/Mouse Transcriptome Isolation Kit) and \$45.00 (NEBNext rRNA Depletion Kit H/M/R). In contrast, we calculate the cost of DASH at less than \$4 per sample when Cas9 and T7 RNA polymerase are made in-house — a very sensible solution for labs that are already spending large amounts of money on NGS. Where Cas9 production is not possible, DASH can still be carried out using commercially available Cas9 protein.

DASH may also enhance the detection of rare mutant alleles that are important for liquid biopsy cancer diagnostics. Allelic depletion with DASH increases the signal (oncogenic mutant allele) to noise (wild-type allele) by more than 60-fold when studying

the KRAS hotspot mutant p.G12D. Other approaches for enriching lowabundance mutations exist, such as restriction enzyme digestion and COLD-PCR. However, these methods are limited when large mutation panels are required. Here we have described a single application for DASH in cancer, but the utility of this method will be fully realized by multiplexing large panels of mutation sites, using guide RNAs and PAM sites as a way to essentially create programmable restriction enzymes that can be used in a single pool. With the rapidly growing number of oncologic therapies that target particular cancer mutations, sensitive and non-invasive techniques for cancer allele detection are increasingly relevant for optimizing patient care.<sup>123</sup> These same techniques are also becoming increasingly important for diagnosis of earlier stage (and generally more curable) cancers as well as the detection of cancer recurrence without needing to re-biopsy the patient.<sup>100,112</sup>

The potential applications of DASH are manifold. Currently, DASH can be customized to deplete any set of defined PAM-adjacent sequences by designing specific libraries of sgRNAs. Given the popularity and promise of CRISPR technologies, we anticipate the adaptation and/or engineering of CRISPR-associated nucleases with more diverse PAM sites.<sup>128,129,132</sup> A portfolio of next-generation Cas9-like nucleases would further enable DASH to deplete large and diverse numbers of arbitrarily selected alleles across the genome without constraint. We envision that DASH will be immediately useful for the development of non-invasive diagnostic tools, with applications to low input samples or cell-free DNA, RNA, or methylation targets in body fluids.<sup>102,104,133,134,135,136,137</sup>



Many other NGS applications could also benefit from depletion of specific sequences, including hemoglobin mRNA depletion for RNA-Seq of blood samples and tRNA depletion for ribosome profiling studies.<sup>138</sup> Depletion of pseudogenes or otherwise homologous sequences by small but consistent differences in sequences is also theoretically possible, and may serve to remove ambiguities in clinical high-throughput sequencing. Using DASH to enrich for minority variations in microbial samples may enable early discovery of pathogen drug resistance. Similarly, the application of DASH to the analysis of cell-free DNA may augment our ability to detect early markers of drug resistance in tumors.<sup>123</sup>

#### **4.5 CONCLUSIONS**

Here, we have demonstrated the broad utility of DASH to enhance molecular signals in diagnostics and its potential to serve as an adaptable tool in basic science research. While the degree of regional depletion of mitochondrial rRNA was sufficient for our application, the depletion parameters were not maximized: we used only 54 sgRNA target sites out of about 250 possible *S. pyogenes* Cas9 sgRNA candidates in the targeted mitochondrial region. Future studies will explore the upper limit of this system while elucidating the most effective sgRNA and CRISPR-associated nuclease selections, which will likely differ based on target and application. Irrespective, depletion of unwanted sequences by DASH is highly generalizable and may effectively lower costs and increase meaningful output across a broad range of sequence-based approaches.

## **4.6 MATERIALS AND METHODS**

### **4.6.1 Generation of cDNA from HeLa cell line and clinical samples**

CSF samples were collected under the approval of the institutional review boards of the University of California San Francisco and San Francisco General Hospital. Samples were processed for high-throughput sequencing as previously described.<sup>75,76</sup> Briefly, amplified cDNAs were made from randomly primed total RNA extracted from 250  $\mu$ L of CSF or 250 pg of HeLa RNA using the NuGEN Ovation v.2 kit (NuGEN, San Carlos, CA, USA) for low nucleic acid content samples. A Nextera protocol (Illumina, San Diego, CA, USA) was used to add on a partial sequencing adapter on both sides.

### **4.6.2 In vitro preparation of the CRISPR/Cas9 complex**

The Cas9 expression vector, containing an N-terminal MBP tag and C-terminal mCherry, was kindly provided by Dr. Jennifer Doudna. The protein was expressed in BL21 Rosetta cells for three hours at 18 °C. Cells were pelleted and frozen. Upon thawing, cells from a 4 L culture preparation were resuspended in 50 mL of lysis buffer (50 mM sodium phosphate pH 6.5, 350 mM NaCl, 1 mM TCEP (tris(2-carboxyethyl)phosphine), 10 % glycerol) supplemented with 0.5 mM EDTA, 1  $\mu$ M PMSF (phenylmethanesulfonyl), and a single Roche complete EDTA-free protease inhibitor tablet (Roche Diagnostics, Indianapolis, IN, USA) and passed through an HC-8000 homogenizer (Microfluidics, Westwood, MA, USA) five times. The lysate was clarified by centrifugation at 20,000 rpm for 45 min at 4 °C and then filtered through a 0.22  $\mu$ m vacuum filtration unit. The filtered lysate was loaded onto three 5 mL HiTrap Heparin HP columns (GE Healthcare, Little Chalfont, UK) arranged in series on a GE AKTA Pure

system. The columns were washed extensively with lysis buffer, and the protein was eluted with a gradient of lysis buffer to buffer B (lysis buffer supplemented with NaCl up to 1.5 M). The resulting fractions were analyzed by Coomassie gel, and those containing Cas9 (centered around the point on the gradient corresponding to 750 mM NaCl) were combined and concentrated down to a volume of 1 mL using 50 K MWCO Amicon Ultra-15 Centrifugal Filter Units (EMD Millipore, Billerica, MA, USA) and then fed through a 0.22 µm syringe filter. Using the AKTA Pure, the 1 mL of filtered protein solution was then injected onto a HiLoad 16/600 Superdex 200 size exclusion column (GE Healthcare, Little Chalfont, UK) pre-equilibrated with buffer C (lysis buffer supplemented with NaCl up to 750 mM). Resulting fractions were again analyzed by Coomassie gel, and those containing purified Cas9 were combined, concentrated, supplemented with glycerol up to a final concentration of 50 %, and frozen at -80 °C until use. Protein concentration was determined by BCA assay. Yield was approximately 80 mg from 4 L of bacterial culture.

sgRNA target sites were selected as described in the main text. DNA templates for sgRNAs based on an optimized scaffold were made with a similar method to that described Lin et al.<sup>139</sup> For each chosen target, a 60mer oligo was purchased including the 18-base T7 transcription start site, the targeted 20mer, and the first 22 bases of the tracr RNA: (5'-TAATACGACTCACTATAGNN  
NNNNNNNNNNNNNNNNNNNGTTTAAGAGCTATG CTGGAAAC-3'). This was mixed with a 90mer representing the 3' end of the sgRNA on the opposite strand (5'-  
AAAAAAAGCACCGACTCGGTGCCACTTTTTTC  
AAGTTGATAACGGACTAGCCTTATTTAACTTGC

TATGCTGTTTCCAGCATAGCTCTTA-3'). DNA templates for T7 sgRNA transcription were then assembled and amplified with a single PCR reaction using primers 5'-TAATACGACTCACTATAG-3' and 5'-AAAAAA AGCACCGACTCGGTGC-3'. The resulting 131 base pair (bp) transcription templates, with the sequence 5'-TAATACGACTCACTATAGNNNNNNNNNNNNNNNNNNNNNNNGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTTT3', were pooled (for the mitochondrial rRNA library), or transcribed separately (for the KRAS experiments). All oligos were purchased from IDT (Integrated DNA Technologies, Coralville, IA, USA).

Transcription was performed using custom-made T7 RNA polymerase (RNAP).<sup>140,141</sup> In each 50  $\mu$ L reaction, 300 ng of DNA template was mixed with T7 RNAP (final concentration 8 ng/ $\mu$ L), buffer (final concentrations of 40 mM Tris pH 8.0, 20 mM MgCl<sub>2</sub>, 5 mM DTT, and 2 mM spermidine), and Ambion brand NTPs (ThermoFisher Scientific, Waltham, MA, USA) (final concentration 1 mM each ATP, CTP, GTP and UTP), and incubated at 37 °C for 4h. Typical yields were 2–20  $\mu$ g of RNA. sgRNAs were purified with a Zymo RNA Clean & Concentrator-5 kit (Zymo Research, Irvine, CA, USA), aliquoted, stored at –80 °C, and used only a single time after thawing.

#### **4.6.3 CRISPR/Cas9 treatment**

To form the ribonucleoprotein (RNP) complex, Cas9 and the sgRNAs were mixed at the desired ratio with Cas9 buffer (final concentrations of 50 mM Tris pH 8.0, 100 mM

NaCl, 10 mM MgCl<sub>2</sub>, and 1 mM TCEP), and incubated at 37 °C for 10 min. This complex was then mixed with the desired amount of sample cDNA in a total of 20 µL, again in the presence of Cas9 buffer, and incubated for 2 h at 37 °C.

Since Cas9 has high nonspecific affinity for DNA it was necessary to disable and remove the Cas9 before continuing.<sup>122</sup> For the rRNA depletion samples, 1 µL (at >600 mAU/mL) of Proteinase K (Qiagen, Hilden, Germany) was added to each sample which was then incubated for an additional 15 min at 37 °C. Samples were then expanded to a volume of 100 µL and purified with three phenol:chloroform:isoamyl alcohol extractions followed by one chloroform extraction in 2 mL Phaselock Heavy tubes (5prime, Hilden, Germany). We added 10 µL of 3 M sodium acetate pH 5.5, 3 µL of linear acrylamide and 226 µL of 100% ethanol to the 100 µL aqueous phase of each sample. Samples were cooled on ice for 30 min. DNA was then pelleted at 4 °C for 45 min, washed once with 70% ethanol, dried at room temperature and resuspended in 10 µL water.

In the case of the KRAS samples, Cas9 was disabled by heating the sample at 95 °C for 15 min in a thermocycler and then removed by purifying the sample with a Zymo DNA Clean & Concentrator-5 kit (Zymo Research, Irvine, CA, USA).

#### **4.6.4 High-throughput sequencing and analysis of sequencing data**

Tagmented samples with and without DASH treatment underwent 10–12 cycles of additional amplification (Kapa Amplification Kit, Kapa Biosystems, Wilmington, MA, USA) with dual-indexing primers. A BluePippin instrument (Sage Science, Beverly, MA, USA) was used to extract DNA between 360 and 540 bp. Sequencing libraries were purified using the Zymo DNA Clean & Concentrator-5 kit and amplified again on an

Opticon qPCR machine (MJ Research, Waltham, MA, USA) using a Kapa Library Amplification Kit until the exponential portion of the quantitative PCR signal was found. Sequencing libraries were then pooled and re-quantified with a ddPCR Library Quantification Kit (Bio-Rad, Hercules, CA, USA). Sequencing was performed on portions of one lane in an Illumina HiSeq 4000 instrument using 135 bp paired-end sequencing.

All reads were quality filtered using PriceSeqFilter (v.1.2)<sup>91</sup> such that only read pairs with less than five ambiguous base calls (defined as Ns or positions with <95% confidence based on Phred score) were retained. Filtered reads were aligned to the hg38 build of the human genome using the STAR aligner (v.2.4.2a).<sup>95</sup> The number of mapped reads per gene and fpkm values were calculated using the exon length and sequence information encoded in the Gencode v.23 primary annotations (GTF file). Library complexity was determined by calculating the reduction in library size after clustering using the cd-hit-dup package.<sup>96</sup> Pathogen-specific alignments to 16S and 18S sequences were accomplished using Bowtie2.<sup>60</sup> Per-nucleotide coverage was calculated from alignment (SAM/BAM) files using the SAMtools suite and analyzed with custom Python scripts utilizing the Pandas data package.<sup>142</sup> Plots were generated with Matplotlib.<sup>143</sup>

#### **4.6.5 ddPCR of KRAS mutant DNA**

KRAS wild-type DNA was obtained from a healthy consenting volunteer. The sample sat until cell separation occurred, and DNA was extracted from the buffy coat with the QIAamp Blood Mini Kit (Qiagen, Hilden, Germany). KRAS G12D genomic DNA

from the human leukemia cell line CCRF-CEM was purchased from ATCC (Manassas, VA, USA). All DNA was sheared to an average of 800 bp using a Covaris M220 (Covaris, Woburn, USA) following the manufacturer's recommended settings. Cas9 reactions occurred as described above.

A primer/probe pair was designed with Primer3 (v2.1)<sup>144</sup> targeting the relatively common KRAS G12D (c.35G>A) mutation. Reactions were thermocycled according to manufacturer protocols using a two-step PCR. An ideal 62 °C annealing/extension temperature was determined by a gradient experiment to ensure proper separation of FAM and HEX signals. The PCR primers and probes used were as follows (purchased from IDT): forward 5'- TAGCTG TATCGTCAAGGCAC-3', reverse 5'-GGCCTGCTGAA AATGACTGA-3'; wild-type probe, 5'-/5HEX/TGCCT ACGC/ZEN/CACAGCTCCA/3IABkFQ/-3'; mutant probe, 5'-/56-FAM/TGCCTACGC/ZEN/CACAGCT CCA/3IABkFQ/-3', with <> denoting the mutant base location, 5HEX and 56-FAM denoting the HEX and FAM reporters, and ZEN and 3IABkFQ denoting the internal and 3' quenchers. Original samples and those subjected to DASH were measured with the ddPCR assay on a Bio-Rad QX100 Droplet Digital PCR system (Bio-Rad, Hercules, CA, USA), following the manufacturer's instructions for droplet generation, PCR amplification, and droplet reading, and using best practices. Pure CCRF-CEM samples were approximately 30% G12D and 70% wild type; all calculations of starting mixtures were made based on this starting ratio.

#### **4.6.6 Ethics**

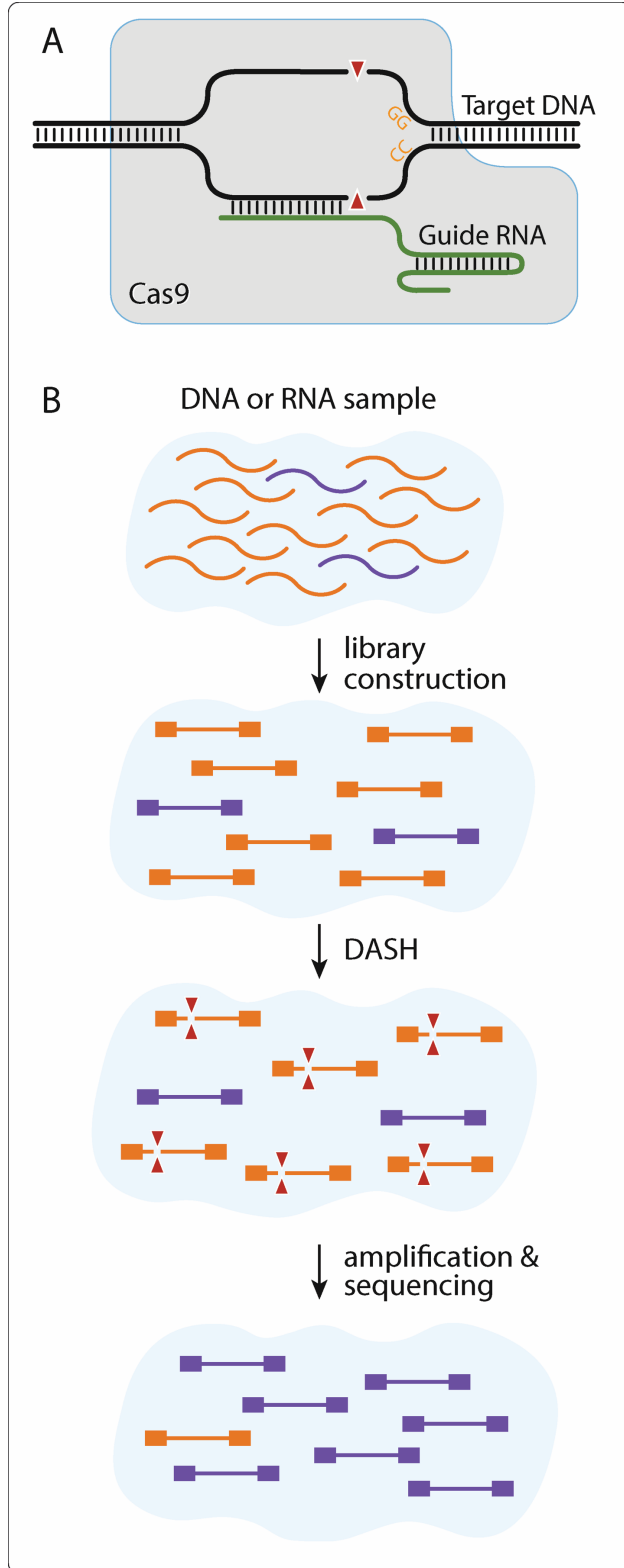
CSF samples, as well as a whole blood sample for the KRAS negative control, were collected under the approval of the institutional review boards of the University of California San Francisco and San Francisco General Hospital (IRB number 13-12236). All experimental methods comply with the Helsinki Declaration.

#### **4.6.7 Availability of data and materials**

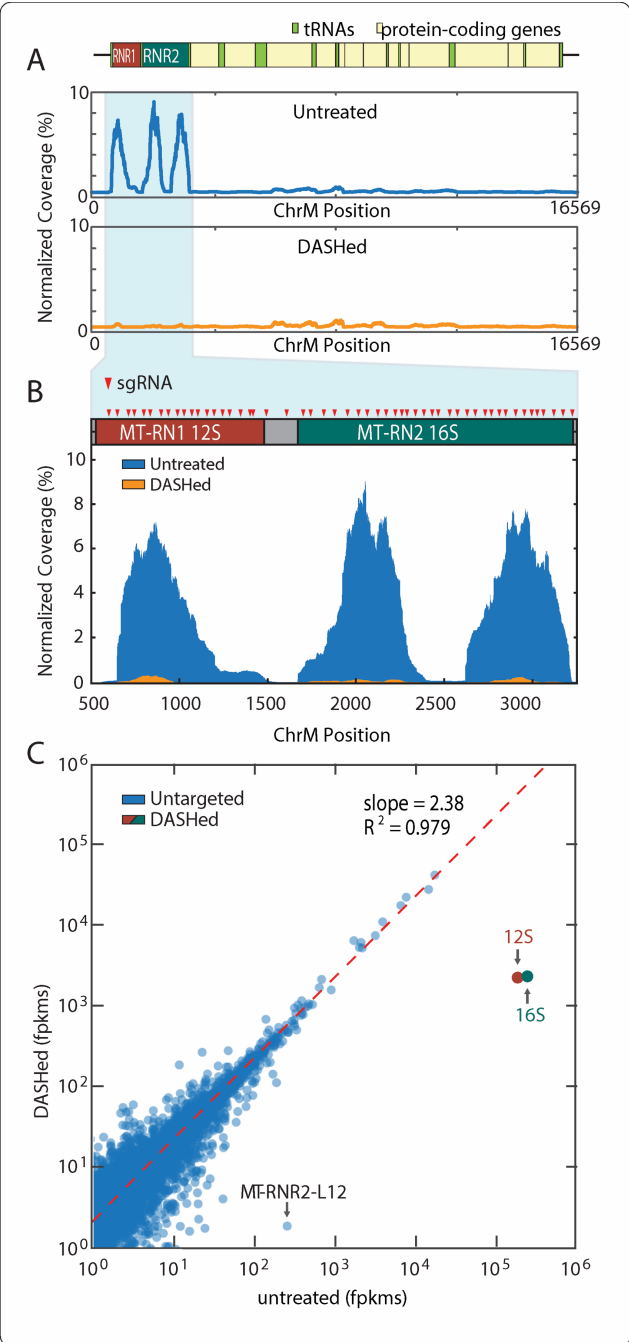
All sequencing data for human subjects has been deposited to NCBI's database of Genotypes and Phenotypes (dbGaP) and can be accessed at <http://www.ncbi.nlm.nih.gov/gap> by entering study accession number phs001067.v1.p1. Sequencing data for HeLa samples has been deposited as a separate BioProject in NCBI's Sequence Read Archive (SRA) and can be found at <http://www.ncbi.nlm.nih.gov/bioproject> by entering study accession number PRJNA311047. Reagents are available upon request from J.L.D



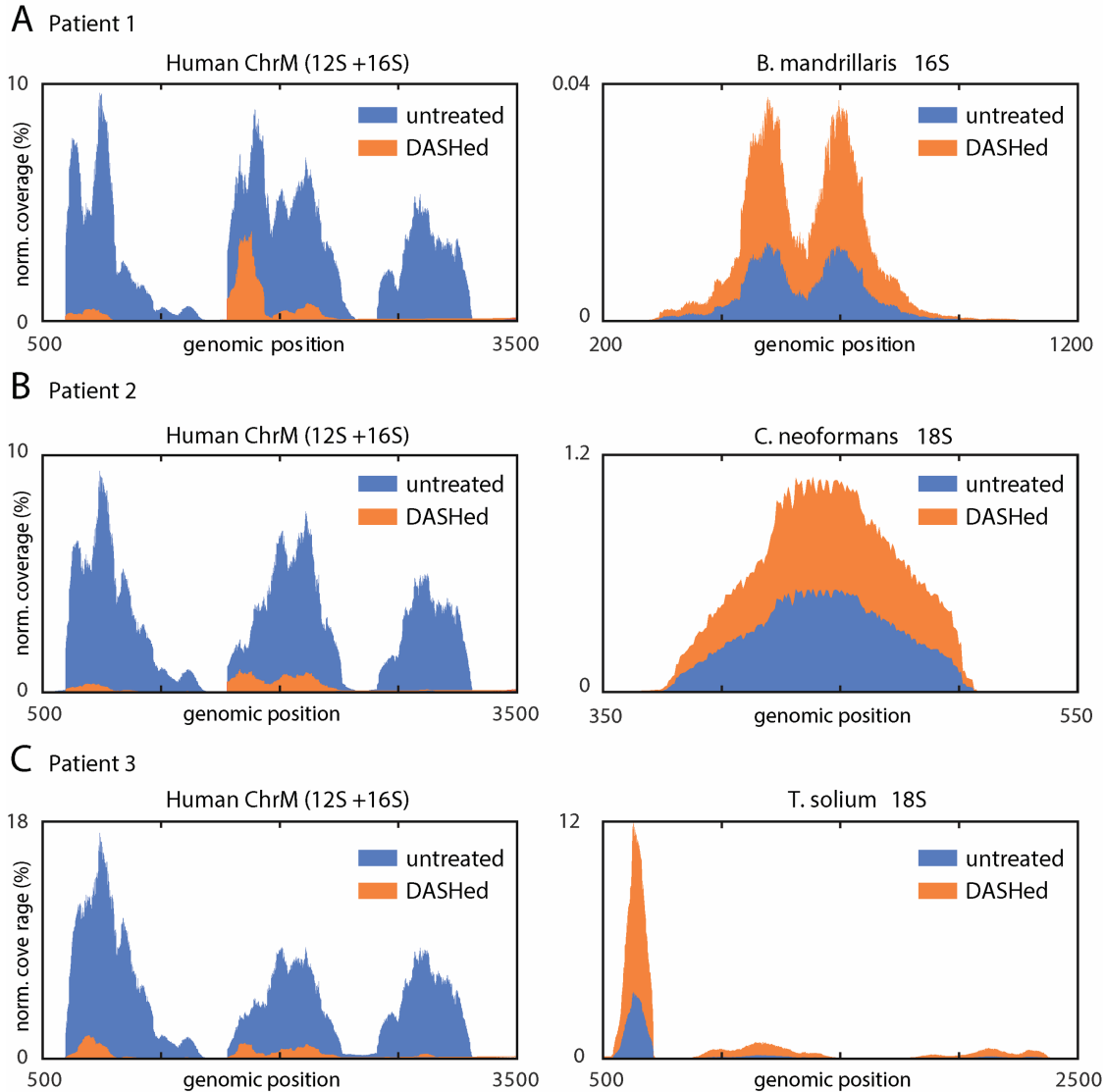
## 4.7 FIGURES



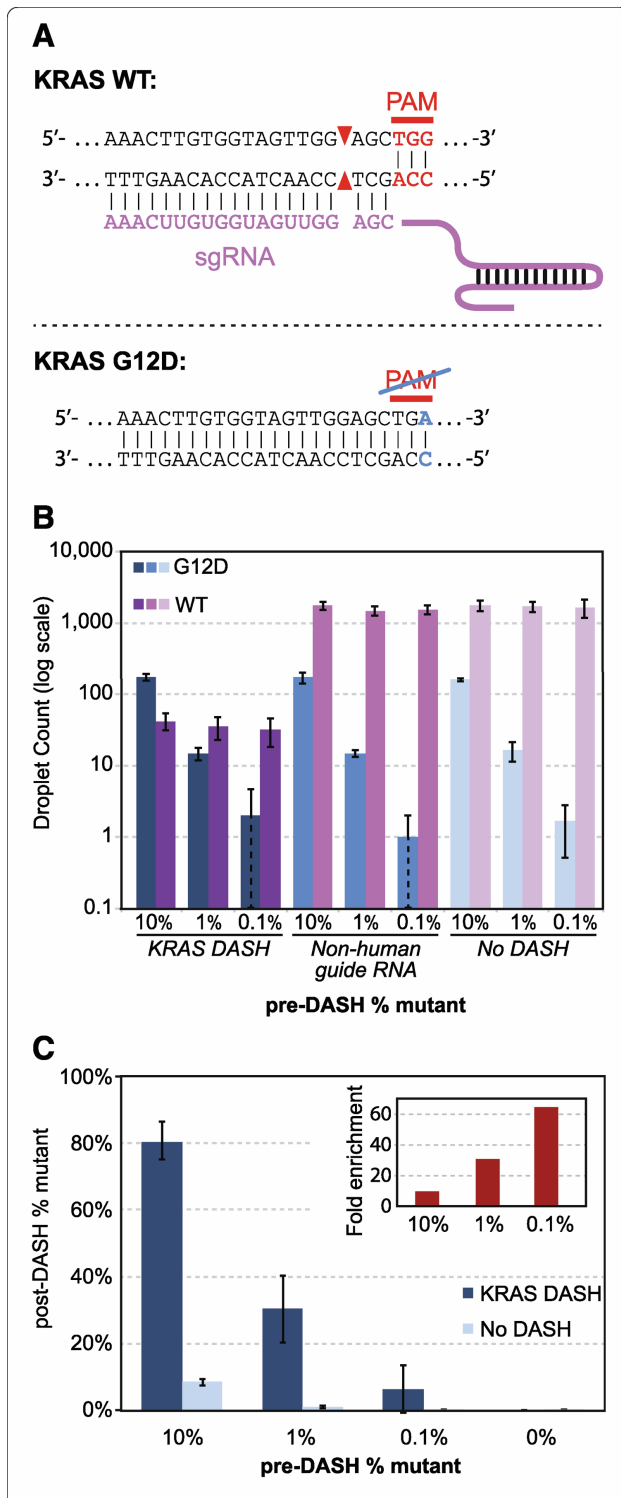
**Figure 4.1 - DASH conceptual figure.** *pyogenes* Cas9 protein binds specifically to DNA targets that match the 'NGG' protospacer adjacent motif (PAM) site. Additional sequence specificity is conferred by a single guide RNA (sgRNA) with a 20-nucleotide hybridization domain. DNA double strand cleavage occurs three nucleotides upstream of the PAM site. b Depletion of Abundant Sequences by Hybridization (DASH) is used to target regions that are present at a disproportionately high copy number in a given next-generation sequencing library following tagmentation or flanking sequencing adaptor placement. Only non-targeted regions that have intact adaptors on both ends of the same molecule are subsequently amplified and represented in the final sequencing library.



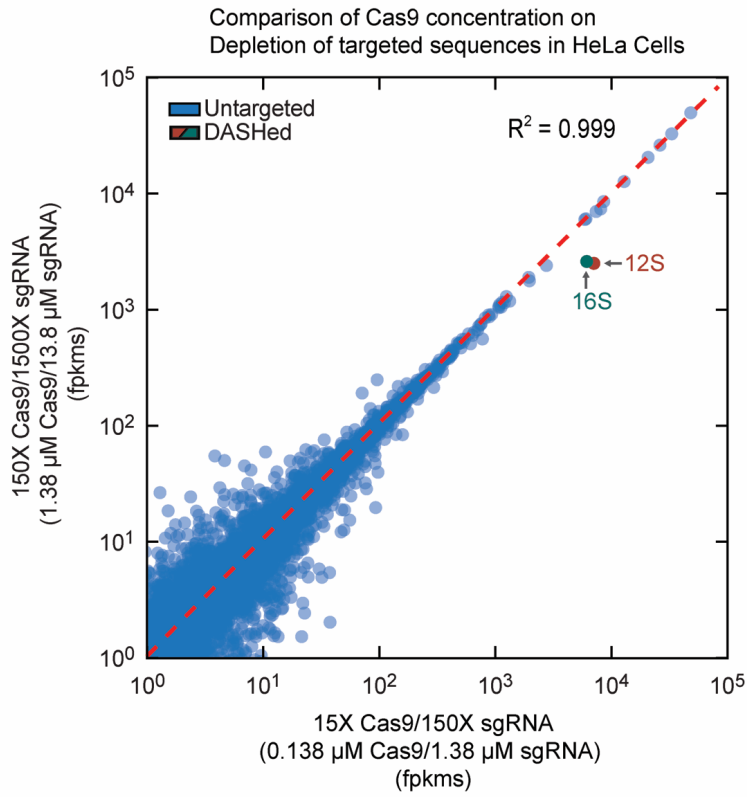
**Figure 4.2 - DASH targeting abundant mitochondrial rRNA in HeLa extractions** A.) Normalized coverage plots showing alignment to the full-length human mitochondrial chromosome. Before treatment, three distinct peaks representing the 12S and 16S ribosomal subunits characteristically account for a large majority of the coverage (>60 % of total mapped reads). After treatment, the peaks are virtually eliminated — with 12S and 16S signatures reduced 1000-fold to 0.055% of mapped reads. B.) Coverage plot of targeted region with 12S and 16S gene boundaries across the top. Red arrowheads represent sgRNA target sites. We chose 54 target sites, spaced approximately 50 bp apart. C.) Scatterplot of the log(fpkm) values per human gene in the control versus treated samples illustrate the significant reduction in reads mapping to the targeted 12S and 16S genes. DASH treatment results in 82 and 105-fold reductions in coverage for the 12S and 16S subunits, respectively. DASH results in 2.38-fold enrichment in reads mapped to untargeted transcripts.  $R^2$  value (0.979) indicates minimal off-target depletion. Between replicates, the R2 coefficient between fpkm values across all genes is 0.994, indicating high reproducibility (three replicates). Notably, one gene, MT-RNR2-L12 (MT-RNR2-like pseudogene), shows significant depletion in the DASHed samples compared with the control



**Figure 4.3 - DASH'ed clinical samples.** Normalized coverage plots of DASH-treated (orange) and untreated (blue) libraries generated from patient cerebrospinal fluid (CSF) samples with confirmed infections. Targeted mitochondrial rRNA genes (left) and representative genes for pathogen diagnosis (right) are depicted for the following: patient 1, *Balamuthia mandrillaris* (a), patient 2, *Cryptococcus neoformans* (b), patient 3, *Taenia solium* (c). Across all cases, the DASH technique significantly reduced the coverage of human 12S and 16S genes by an average of 7.5-fold while increasing the coverage depth for pathogenic sequences by an average 5.9-fold. See Table 4.1 for relevant data.



**Figure 4.4 - DASH in cancer.** A.) DASH is used to selectively deplete one allele while keeping the other intact. An sgRNA in conjunction with Cas9 targets a wild-type (WT) KRAS sequence. However, since the G12D (c.35G>A) mutation disrupts the PAM site, Cas9 does not efficiently cleave the mutant KRAS sequence. Subsequent amplification of all alleles using flanking primers, as in the case of digital PCR, Sanger sequencing, or high-throughput sequencing, is only effective for non-cleaved and mutant sites. B.) Three human genomic DNA samples with varying ratios of wild-type to mutant (G12D) KRAS were treated either with KRAS-targeted DASH, a non-human control DASH, or no DASH. Counts of intact wild-type and G12D sequences were then measured by droplet digital PCR (ddPCR). C.) Same data as in (B), presented as percentage of mutant sequences detected. Inset shows fold enrichment of the percentage of mutant sequences with KRAS-targeted DASH versus no DASH. For both (B) and (C), values and error bars are the average and standard deviation, respectively, of three independent experiments.



**Figure 4.5 - Concentration-dependent depletion.** Scatterplot of log of fragments per kilobase of transcript per million mapped reads (log-fpkms) values per human gene comparing HeLa cells DASHed with two different Cas9/sgRNA concentrations.

mt-rRNA 16S AAGTGCACCTGGACGAA**CCA**GAGTGTAGCTTAAACACAAAGCACCCAACTTACACTTAGGAGATTTCAACTTAACTTGACCGCTCTGAGCTAAACCTAGC 100  
 mtRNR2L 12 AAGTGCACCTGGACGAA**CCA**GAGTGTAGCTTAAACATAAAGCACCCAACTTACACTTAGGAGATTTCAACT**CA**ACTTGAC**CA**CTCTGAG**CA**AACCTAGC 100

mt-rRNA 16S CCCAAACCACCT**CCA**CCTTACTACCAGACAACCTTAGCCAAACATTACCCAAATAAAGTAT**AGG**CGATAGAAATTGAAACCTGGCGCAATAGATATAG 200  
 mtRNR2L 12 CCTAAACC**CGTTC**CACTTACTATCA**AA**TAACTTAA**CCAA**ACCATTACCCAAATAAAGTAT**AGG**CGATAGAAATT**GTAA**ACC**GGCG**CAATAGATATAG 200

mt-rRNA 16S TACCGCAAGGAAAGATGAAAAATTATA**CCA**AGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAATTAACTAGAAA**TA**CTTTGC**AA** 300  
 mtRNR2L 12 TACCGCAAGGAAAGATGAAAAATTATA**CCA**AGCATAAT**ACAG**CAAGGACTAACCCCT**GT**ACCTTTGCATAATGAATTAACTAGAAA**TA**CTTTGC**AA** 300

mt-rRNA 16S **GG**AGAGCCAAAGCTAAGACCCCGAAACCAGACGAGCTACCTAAGAACAGCTAAAAGAGCACA**CC**CGTCTATGTAGCAAAATAGTGGGAAGATTTATAGG 400  
 mtRNR2L 12 **AG**AGAA**CCAA**AGCTAAG**CG**CCCGAAACCAGACGAGCTACCTAAGAACAGCTAAAAGAGCACA**CC**CGTCTATGTAGCAAAATAGTGGGAAGATTTATAGG 400

mt-rRNA 16S TAGAGGCGACAAACCT**ACC**GAGCCTGGTGATAGCT**GG**TTGTCCAAGATAGAATCTTAGTTCAACTTTAAATTTGC**CCA**CAGAACCCTCTAAATCCCC**TT**G 500  
 mtRNR2L 12 TAGAGGCGACAAACCT**AT**CGAGCCTGGTGATAGCT**GG**TTGTCCAAGATAGAATCTTAGTTCAACTTTAAATTT**AC**CT**AC**AGAACC**TTCT**AAATCCCC**TT**G 500

mt-rRNA 16S TAAATTTAACTGTTAGTCCAAAGAGGAACAGCTCTTTGGACACT**AG**GAAAAACCTTGTAGAGAGAGTAAAAATTTAAAC**CCA**TAGTAGGCC**TTAAA**AG 600  
 mtRNR2L 12 TAAATTTAACTGTTAGTCCAAAGAGGAACAGCTCTTTGGACACT**AG**GAAAAACCTTGTAAAGAGAGTAAAAATTTAA**TACCA**TAGTAGGCC**TTAAA**AG 600

mt-rRNA 16S CAGCCACCAATTAAGAAAGCGTTCAAGCTCAACA**CC**CACTACCTAAAAATCCCAAACATATAACTGAACT**CCT**CACACCCAAATGGACCAATCTATCAC 700  
 mtRNR2L 12 CAGCCACCAATTAAGAAAGCGTTCAAGCTCAACA**CC**CA**TC**CGTCTAAAAATCCCAAACAT**CA**ACTGAG**CTCCT**TACACTCAATGGACCAATCTAT**TAC** 700

mt-rRNA 16S **CCT**ATAGAAGAACTAATGTTAGTATAAGTAACATGAAACATTTCTCCT**CG**CATAAGCCTGCGTCAGATCAAAACACTGAACTGACAATTAACAG**CCCA** 800  
 mtRNR2L 12 **CTT**ATAGAAGAACTAATGTTAGTATAAGTAACATGAAACATTTCTCCT**CG**CATAAGCCT**ACA**TCAGAC**CCAAA**T**ATT**AACTGACAATTAACAG**CCCA** 800

mt-rRNA 16S TATCTACAATCAACCAACAAGTCATTATTACCTCACTGTCAA**CC**AACACAGGCATGCTCATAAGGAAAGTTAAAAAAGTAAA**AG**GAACTCGGCAAA 900  
 mtRNR2L 12 TATCTACAATCAACCAACAAG**CC**ATTATTACCTCACTGT**TA**CC**CA**AACACAGGCAT**GC**CA**CA**AGGAAAGTTAAAAAAGTAAA**AG**GAACTCGGCAAA 900

mt-rRNA 16S CCTTACCCCGCTGTTTACCAAAAACATCACCTCTAGCATCA**CCA**GTATTAGAGGCACCGCTGCCAGTGACACATGTTAACGG**CCG**CGGTACCCTAA 1000  
 mtRNR2L 12 **TCT**TACCCCGCTGTTTACCAAAAACATCACCTCTAGCAT**TAT**CAGTATTAGAGGCACCGCTGCC**CG**GTGACAT**AT**GTTTAA**CGG**CC**GGT**ACCCTAA 1000

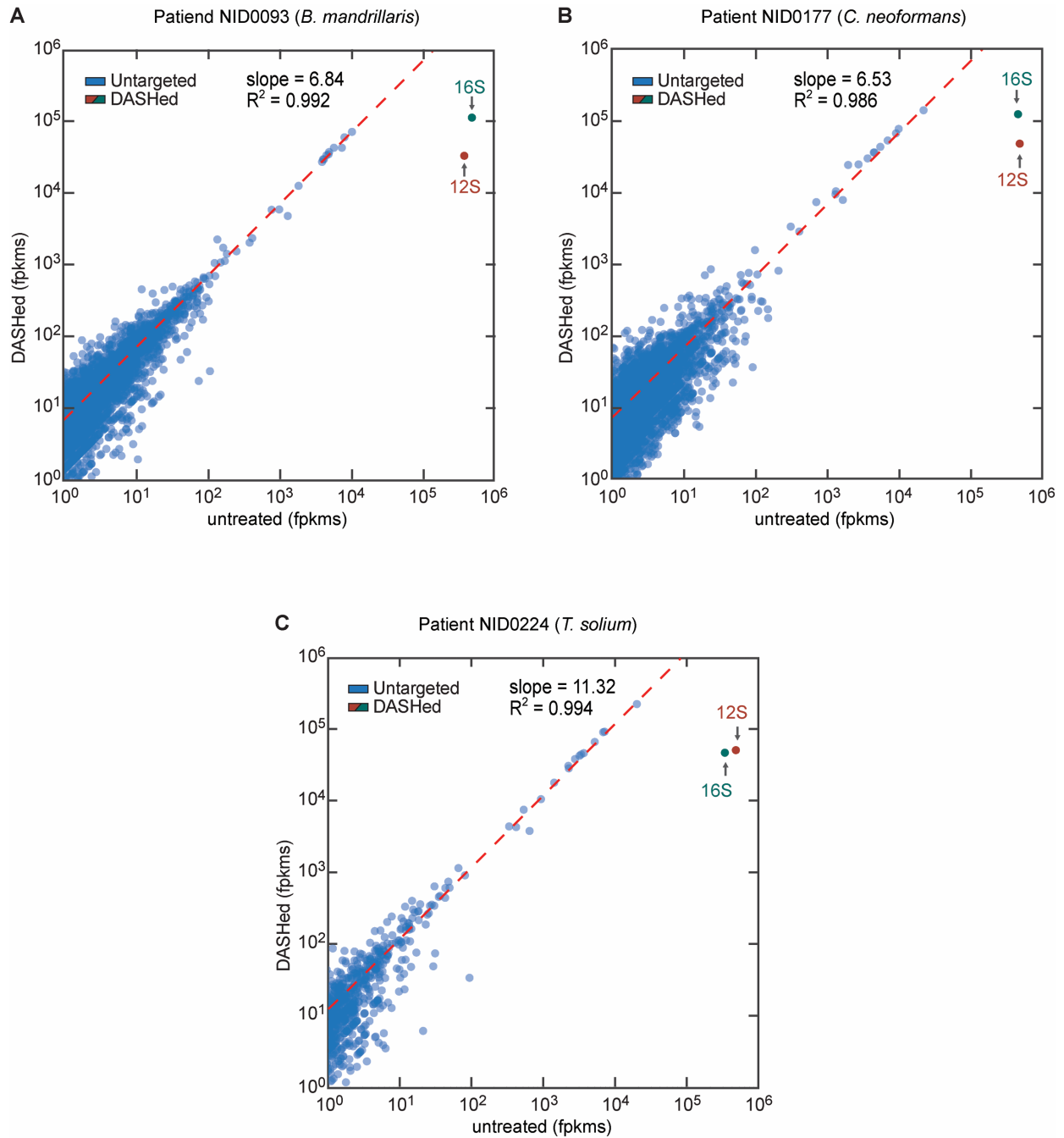
mt-rRNA 16S CCGTGCAAAGGTAGCATAATCACTTGTCTTAAATAGGG**ACT**GTATGAATGGCTCCACGAGGGTT**CAG**CTGTCTCTTACTTTTAA**CCA**GTGAAATTGA 1100  
 mtRNR2L 12 CCGTGCAA**AG**GTAGCATAATCACTTGTCTTAAATAGGG**ACT**GTATGAATGGCTCCACGAGGGTT**CAG**CTGTCTCTTACTTT**CA**AC**CA**GTGAAATTGA 1100

mt-rRNA 16S CCTGCCCGTGAAGAGGCGGGCATGACACAGCAAGACGAGAAGA**CC**TATGGAGCTTTAATTTAATGCAACAGTA**CCT**AACAAACCACAGGTCCTA 1200  
 mtRNR2L 12 CCT**AC**CCCGTGAAGAGGCGGGCAT**ACA**TAGCAAGACT**TAT**GTACT**ACA**T**ACA**ACT**AAC**ATGTT**TACT**GAGCAAG**TTTT**CTCAAAAAGT**AG**AAT**GCTAG**G 1200

mt-rRNA 16S AACTACCAACCTGCATTAAAAATTT**CG**GT**GG**CGAC  
 mtRNR2L 12 **ATT**ATAGAAA**ATA**-ATTAAAA**TAA**T**ACA**TTT**CAC**CC**TT**

Key to mtRNR2L 12 sequence: **PAM**  
 No PAM  
 match  
 mismatch  
 sgRNA target, perfect match  
 sgRNA target, some mismatches

Figure 4.6 - Mitochondrial rRNA target sites in mtRNR2L12



**Figure 4.7 - Patient-specific scatterplots.** Scatterplots of the log of fragments per kilobase of transcript per million mapped reads (log-fpkms) values per human gene in the DASHed vs. untreated patient samples. The slopes of the regression lines (red) indicate the fold enrichment in reads mapped to untargeted transcripts. R-squared (R<sup>2</sup>) values of the regression lines indicate minimal off-target depletion.

## 4.8 TABLES

**Table 4.1 - Summary of depletion/enrichment results in DASH-treated CSF**

Pathogen	Read count (percentage duplicates)		Targeted genes (fpkm)				Representative pathogenic gene <sup>d</sup> (fold change)		R <sup>2</sup> non-targeted genes, untreated versus DASHed
			12S		16S				
	Un-treated	DASHed	Un-treated	DASHed	Un-treated	DASHed	Un-treated	DASHed	
B. mandrillaris	1.81 M (26 %)	2.54 M (15 %)	298,922	28,005	380,073	93,164	0.028 %	0.102 %	0.992
							(3.6x)		
C. neoformans	2.95 M (27 %)	3.43 M (11 %)	361,501	37,168	342,857	93,703	1.5 %	15.4 %	0.986
							(10.3x)		
T. solium	2.38 M (33 %)	1.89 M (30 %)	451,044	46,993	317,640	43,257	12.0 %	44.3 %	0.994
							(3.7x)		

<sup>a</sup> Representative genes are 16S for *Balamuthia mandrillaris* and 18S for *Cryptococcus neoformans* and *Taenia solium*

**Table 4.2 - sgRNA targeted sequences**

Name	Target Sequence	Name	Target Sequence
mt-rRNA-1	ATTTTCAGTGTATTGCTTTG	mt-rRNA-29	GGAACAGCTCTTTGGACACT
mt-rRNA-2	ACATCACCCATAAACAAAT	mt-rRNA-30	GGCTGCTTTTAGGCCTACTA
mt-rRNA-3	AGGGTGAACACTGGAACG	mt-rRNA-31	TTTGGGATTTTTAGGTAGT
mt-rRNA-4	TCTAAATCACACGATCAAA	mt-rRNA-32	GATTGGTCCAATTGGGTGTG
mt-rRNA-5	TTTCCCGTGGGGGTGTGGCT	mt-rRNA-33	ACTAACATTAGTTCCTCTAT
mt-rRNA-6	AAACTTTCGTTTATTGCTAA	mt-rRNA-34	TGATCTGACGCAGGCTTATG
mt-rRNA-7	AATCGTGTGACCGCGGTGGC	mt-rRNA-35	TGTTGGTTGATTGTAGATAT
mt-rRNA-8	ATCTAAAACACTCTTTACGC	mt-rRNA-36	CTTATGAGCATGCCTGTGTT
mt-rRNA-9	ACTGGAGTTTTTACAACCTC	mt-rRNA-37	GAAAGGTTAAAAAAGTAAA
mt-rRNA-10	CACAAAATAGACTACGAAAG	mt-rRNA-38	GCAGGCGGTGCCTCTAATAC
mt-rRNA-11	GGGGTATCTAATCCCAGTTT	mt-rRNA-39	TTTGCACGGTTAGGGTACCG
mt-rRNA-12	GATTAACTGTTGAGGTTTA	mt-rRNA-40	CCTCGTGGAGCCATTTCATC
mt-rRNA-13	GTCCTTTGAGTTTTAAGCTG	mt-rRNA-41	CACGGGCAGGTCAATTCAC
mt-rRNA-14	ACAGAACAGGCTCCTCTAGA	mt-rRNA-42	TAATAAATTAAGCTCCATA
mt-rRNA-15	TATATAGGCTGAGCAAGAGG	mt-rRNA-43	TTAGGACCTGTGGGTTTGTT
mt-rRNA-16	TCTTCAGCAAACCCTGATGA	mt-rRNA-44	TGCATTAAAAATTTTCGGTTG
mt-rRNA-17	CCCATTCTTGCCACCTCAT	mt-rRNA-45	AAGTCTTAGCATGTACTGCT
mt-rRNA-18	TCGACCCTTAAGTTTCATAA	mt-rRNA-46	TGTTCCGTTGGTCAAGTTAT
mt-rRNA-19	TGAAACTTAAGGGTCGAAGG	mt-rRNA-47	GTTGATATGGA CTCTAGAAT
mt-rRNA-20	GTATACTTGAGGAGGGTGAC	mt-rRNA-48	TACGACCTCGATGTTGGATC
mt-rRNA-21	CTTTGTGTTAAGCTACACTC	mt-rRNA-49	GATGGTGCAGCCGCTATTAA
mt-rRNA-22	AAGGTTGTCTGGTAGTAAGG	mt-rRNA-50	GGTCTGAACTCAGATCACGT
mt-rRNA-23	CATTTACCCAAATAAAGTAT	mt-rRNA-51	TCTTGTCTTTTCGTACAGGG
mt-rRNA-24	AGTCCTTGCTATATTATGCT	mt-rRNA-52	TGAGATGATATCATTTACGG
mt-rRNA-25	TAAC TAGAAATAACTTTGCA	mt-rRNA-53	CCCACACCCACCCAAGAACA
mt-rRNA-26	CACTATTTTGCTACATAGAC	mt-rRNA-54	ACTTAAAAC TTTACAGTCAG
mt-rRNA-27	CTACCGAGCCTGGTGATAGC	KRAS WT	AAACTTGTGGTAGTTGGAGC
mt-rRNA-28	AGGGGATTTAGAGGGTTCTG	Non-human control	ACAAATATTTTAATACATGA



## REFERENCES

1. Vora, N. M. *et al.* Burden of encephalitis-associated hospitalizations in the United States, 1998-2010. *Neurology* **82**, 443–451 (2014).
2. Dubey, D. *et al.* Autoimmune encephalitis epidemiology and a comparison to infectious encephalitis. *Ann. Neurol.* **83**, 166–177 (2018).
3. Behjati, S. & Tarpey, P. S. What is next generation sequencing? *Arch Dis Child Educ Pract Ed* **98**, 236–238 (2013).
4. Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLOS Biology* **13**, e1002195 (2015).
5. Zhu, L. *et al.* Diagnosis for choroideremia in a large Chinese pedigree by next-generation sequencing (NGS) and non-invasive prenatal testing (NIPT). *Mol Med Rep* **15**, 1157–1164 (2017).
6. Strom, S. P. Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer Biol Med* **13**, 3–11 (2016).
7. Kaneko, M. *et al.* Identification of vaccine-derived rotavirus strains in children with acute gastroenteritis in Japan, 2012-2015. *PLoS ONE* **12**, e0184067 (2017).
8. He, T., Kaplan, S., Kamboj, M. & Tang, Y.-W. Laboratory Diagnosis of Central Nervous System Infection. *Curr Infect Dis Rep* **18**, 35 (2016).
9. Dando, S. J. *et al.* Pathogens Penetrating the Central Nervous System: Infection Pathways and the Cellular and Molecular Mechanisms of Invasion. *Clinical Microbiology Reviews* **27**, 691–726 (2014).

10. Branton, W. G. *et al.* Brain microbiota disruption within inflammatory demyelinating lesions in multiple sclerosis. *Sci Rep* **6**, 37344 (2016).
11. Pignolet, B. S., Gebauer, C. M. & Liblau, R. S. Immunopathogenesis of paraneoplastic neurological syndromes associated with anti-Hu antibodies: A beneficial antitumor immune response going awry. *Oncol Immunology* **2**, e27384 (2013).
12. Toothaker, T. B. & Rubin, M. Paraneoplastic neurological syndromes: a review. *Neurologist* **15**, 21–33 (2009).
13. Rosenfeld, M. R. & Dalmau, J. Paraneoplastic Neurologic Disorders: A Brief Overview. *Memo* **5**, 197–200 (2012).
14. Pittock, S. J., Kryzer, T. J. & Lennon, V. A. Paraneoplastic antibodies coexist and predict cancer, not neurological syndrome. *Ann. Neurol.* **56**, 715–719 (2004).
15. Horta, E. S. *et al.* Neural autoantibody clusters aid diagnosis of cancer. *Clin. Cancer Res.* **20**, 3862–3869 (2014).
16. O'Brien, T. J., Pasaliaris, B., D'Apice, A. & Byrne, E. Anti-Yo positive paraneoplastic cerebellar degeneration: a report of three cases and review of the literature. *J Clin Neurosci* **2**, 316–320 (1995).
17. Hasadsri, L., Lee, J., Wang, B. H., Yekkirala, L. & Wang, M. Anti-Yo Associated Paraneoplastic Cerebellar Degeneration in a Man with Large Cell Cancer of the Lung. *Case Reports in Neurological Medicine* **2013**, 1–5 (2013).
18. O'Donovan, K. J., Diedler, J., Couture, G. C., Fak, J. J. & Darnell, R. B. The Onconeural Antigen cdr2 Is a Novel APC/C Target that Acts in Mitosis to Regulate C-Myc Target Genes in Mammalian Tumor Cells. *PLoS ONE* **5**, e10045 (2010).

19. Albert, M. L. *et al.* Tumor-specific killer cells in paraneoplastic cerebellar degeneration. *Nature Medicine* **4**, 1321–1324 (1998).
20. Venkatraman, A. & Opal, P. Paraneoplastic cerebellar degeneration with anti-Yo antibodies - a review. *Annals of Clinical and Translational Neurology* **3**, 655–663 (2016).
21. Senties-Madrid, H. & Vega-Boada, F. Paraneoplastic syndromes associated with anti-Hu antibodies. *Isr. Med. Assoc. J.* **3**, 94–103 (2001).
22. Pascale, A., Amadio, M. & Quattrone, A. Defining a neuron: neuronal ELAV proteins. *Cell. Mol. Life Sci.* **65**, 128–140 (2008).
23. Lancaster, E. The Diagnosis and Treatment of Autoimmune Encephalitis. *J Clin Neurol* **12**, 1–13 (2016).
24. Beghetto, E. & Gargano, N. Antigen discovery using whole-genome phage display libraries. *Methods Mol. Biol.* **1061**, 79–95 (2013).
25. Burritt, J. B., Quinn, M. T., Jutila, M. A., Bond, C. W. & Jesaitis, A. J. Topological mapping of neutrophil cytochrome b epitopes with phage-display libraries. *J. Biol. Chem.* **270**, 16974–16980 (1995).
26. Frei, J. C. & Lai, J. R. Protein and Antibody Engineering by Phage Display. *Meth. Enzymol.* **580**, 45–87 (2016).
27. Larman, H. B. *et al.* Autoantigen discovery with a synthetic human peptidome. *Nat. Biotechnol.* **29**, 535–541 (2011).
28. Larman, H. B. *et al.* PhIP-Seq characterization of autoantibodies from patients with multiple sclerosis, type 1 diabetes and rheumatoid arthritis. *J. Autoimmun.* **43**, 1–9 (2013).

29. Graus, F. *et al.* A clinical approach to diagnosis of autoimmune encephalitis. *Lancet Neurol* **15**, 391–404 (2016).
30. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
31. Genetic analysis of pathogenic bacteria: a laboratory manual. Available at: <https://www.cabdirect.org/cabdirect/abstract/19972201503>. (Accessed: 14th November 2018)
32. Xu, G. J. *et al.* Comprehensive serological profiling of human populations using a synthetic human virome. *Science* **348**, aaa0698 (2015).
33. Mohan, D. *et al.* PhIP-Seq characterization of serum antibodies using oligonucleotide-encoded peptidomes. *Nature Protocols* **13**, 1958 (2018).
34. Yuan, T. *et al.* Improved Analysis of Phage ImmunoPrecipitation Sequencing (PhIP-Seq) Data Using a Z-score Algorithm. (2018). doi:10.1101/285916
35. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547–1548 (2008).
36. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing on JSTOR. Available at: [https://www.jstor.org/stable/2346101?seq=1#metadata\\_info\\_tab\\_contents](https://www.jstor.org/stable/2346101?seq=1#metadata_info_tab_contents). (Accessed: 14th November 2018)
37. Emini, E. A., Hughes, J. V., Perlow, D. S. & Boger, J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol* **55**, 836–839 (1985).

38. Frith, M. C., Saunders, N. F. W., Kobe, B. & Bailey, T. L. Discovering Sequence Motifs with Arbitrary Insertions and Deletions. *PLoS Computational Biology* **4**, e1000071 (2008).
39. Gadoth, A. *et al.* Microtubule-associated protein 1B: Novel paraneoplastic biomarker. *Ann. Neurol.* **81**, 266–277 (2017).
40. Berto, S., Usui, N., Konopka, G. & Fogel, B. L. ELAVL2-regulated transcriptional and splicing networks in human neurons link neurodevelopment and autism. *Hum. Mol. Genet.* **25**, 2451–2464 (2016).
41. Wang, H., Molfenter, J., Zhu, H. & Lou, H. Promotion of exon 6 inclusion in HuD pre-mRNA by Hu protein family members. *Nucleic Acids Res* **38**, 3760–3770 (2010).
42. Rousseau, A. *et al.* T cell response to Hu-D peptides in patients with anti-Hu syndrome. *Journal of Neuro-Oncology* **71**, 231–236 (2005).
43. Roberts, W. K. *et al.* Patients with lung cancer and paraneoplastic Hu syndrome harbor HuD-specific type 2 CD8+ T cells. *Journal of Clinical Investigation* (2009).  
doi:10.1172/JCI36131
44. Klein, L., Klugmann, M., Nave, K.-A., Tuohy, V. K. & Kyewski, B. Shaping of the autoreactive T-cell repertoire by a splice variant of self protein expressed in thymic epithelial cells. *Nature Medicine* **6**, 56–61 (2000).
45. Zaharieva, E., Hausmann, I. U., Bräuer, U. & Soller, M. Concentration and localization of co-expressed ELAV/Hu proteins control specificity of mRNA processing. *Molecular and Cellular Biology* MCB.00473-15 (2015). doi:10.1128/MCB.00473-15

46. Single-cell RNA-sequencing resolves self-antigen expression during mTEC development | Scientific Reports. Available at: <https://www.nature.com/articles/s41598-017-19100-4>. (Accessed: 14th November 2018)
47. Földy, C. *et al.* Single-cell RNAseq reveals cell adhesion molecule profiles in electrophysiologically defined neurons. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E5222-5231 (2016).
48. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nature Biotechnology* (2018). doi:10.1038/nbt.4259
49. Honnorat, J. *et al.* Onco-neural antibodies and tumour type determine survival and neurological symptoms in paraneoplastic neurological syndromes with Hu or CV2/CRMP5 antibodies. *J. Neurol. Neurosurg. Psychiatry* **80**, 412–416 (2009).
50. Yu, Z. *et al.* CRMP-5 neuronal autoantibody: marker of lung cancer and thymoma-related autoimmunity. *Ann. Neurol.* **49**, 146–154 (2001).
51. Titulaer, M. J. *et al.* SOX antibodies in small-cell lung cancer and Lambert-Eaton myasthenic syndrome: frequency and relation with survival. *J. Clin. Oncol.* **27**, 4260–4267 (2009).
52. Kazarian, M. & Laird-Offringa, I. A. Small-cell lung cancer-associated autoantibodies: potential applications to cancer diagnosis, early detection, and therapy. *Molecular Cancer* **10**, 33 (2011).
53. Fritzler, M. J., Zhang, M., Stinton, L. M. & Rattner, J. B. Spectrum of centrosome autoantibodies in childhood varicella and post-varicella acute cerebellar ataxia. *BMC Pediatrics* **3**, (2003).

54. Bao, L., Varden, C. E., Zimmer, W. E. & Balczon, R. Localization of autoepitopes on the PCM-1 autoantigen using scleroderma sera with autoantibodies against the centrosome. *Mol Biol Rep* **25**, 111–119 (1998).
55. Eichler, T. W. *et al.* CDR2L Antibodies: A New Player in Paraneoplastic Cerebellar Degeneration. *PLoS ONE* **8**, e66002 (2013).
56. Blachère, N. E. *et al.* T cells targeting a neuronal paraneoplastic antigen mediate tumor rejection and trigger CNS autoimmunity with humoral activation. *Eur. J. Immunol.* **44**, 3240–3251 (2014).
57. Grabar, P. 'Self' and 'not-self' in immunology. *Lancet* **1**, 1320–1322 (1974).
58. Bonilla, N. *et al.* Phage on tap—a quick and efficient protocol for the preparation of bacteriophage laboratory stocks. *PeerJ* **4**, e2261 (2016).
59. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
60. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
61. MEME Suite: tools for motif discovery and searching | Nucleic Acids Research | Oxford Academic. Available at: [https://academic.oup.com/nar/article/37/suppl\\_2/W202/1135092](https://academic.oup.com/nar/article/37/suppl_2/W202/1135092). (Accessed: 19th November 2018)
62. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* **43**, D405–D412 (2015).

63. Thu, K. L. *et al.* SOX15 and other SOX family members are important mediators of tumorigenesis in multiple cancer types. *Oncoscience* **1**, 326–335 (2014).
64. Kerasnoudis, A. Isolated ZIC4 Antibodies in Paraneoplastic Cerebellar Syndrome With an Underlying Ovarian Tumor. *Archives of Neurology* **68**, 1073 (2011).
65. Sabater, L. *et al.* ZIC antibodies in paraneoplastic cerebellar degeneration and small cell lung cancer. *J Neuroimmunol* **201–202**, 163–165 (2008).
66. Han, W. *et al.* Clinicopathologic and Prognostic Significance of the Zinc Finger of the Cerebellum Family in Invasive Breast Cancer. *J Breast Cancer* **21**, 51–61 (2018).
67. Hoang-Minh, L. B. *et al.* PCM1 Depletion Inhibits Glioblastoma Cell Ciliogenesis and Increases Cell Death and Sensitivity to Temozolomide. *Transl Oncol* **9**, 392–402 (2016).
68. Li, L. Y. *et al.* Genetic Profiles Associated with Chemoresistance in Patient-Derived Xenograft Models of Ovarian Cancer. *Cancer Res Treat* (2018). doi:10.4143/crt.2018.405
69. Shiiro, Y. *et al.* Identification and characterization of SAP25, a novel component of the mSin3 corepressor complex. *Mol. Cell. Biol.* **26**, 1386–1397 (2006).
70. Zhao, H. *et al.* Endothelial Robo4 suppresses breast cancer growth and metastasis through regulation of tumor angiogenesis. *Mol Oncol* **10**, 272–281 (2016).
71. Li, Y. *et al.* Expression of Robo protein in bladder cancer tissues and its effect on the growth of cancer cells by blocking Robo protein. *Int J Clin Exp Pathol* **8**, 9932–9940 (2015).
72. Cai, H. *et al.* Overexpression of Roundabout4 predicts poor prognosis of primary glioma patients via correlating with microvessel density. *J. Neurooncol.* **123**, 161–169 (2015).
73. Zunt, J. R. & Baldwin, K. J. Chronic and subacute meningitis. *Continuum (Minneapolis)* **18**, 1290–1318 (2012).



74. Baldwin, K. J. & Zunt, J. R. Evaluation and Treatment of Chronic Meningitis. *Neurohospitalist* **4**, 185–195 (2014).
75. Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing | NEJM. Available at: <https://www.nejm.org/doi/full/10.1056/NEJMoa1401268>. (Accessed: 10th December 2018)
76. Wilson, M. R. *et al.* Diagnosing Balamuthia mandrillaris Encephalitis With Metagenomic Deep Sequencing. *Ann. Neurol.* **78**, 722–730 (2015).
77. Wilson, M. R. *et al.* Acute West Nile Virus Meningoencephalitis Diagnosed Via Metagenomic Deep Sequencing of Cerebrospinal Fluid in a Renal Transplant Patient. *Am. J. Transplant.* **17**, 803–808 (2017).
78. Murkey, J. A. *et al.* Hepatitis E Virus-Associated Meningoencephalitis in a Lung Transplant Recipient Diagnosed by Clinical Metagenomic Sequencing. *Open Forum Infect Dis* **4**, ofx121 (2017).
79. Wilson, M. R. *et al.* A novel cause of chronic viral meningoencephalitis: Cache Valley virus. *Ann. Neurol.* **82**, 105–114 (2017).
80. Naccache, S. N. *et al.* Diagnosis of neuroinvasive astrovirus infection in an immunocompromised adult with encephalitis by unbiased next-generation sequencing. *Clin. Infect. Dis.* **60**, 919–923 (2015).
81. Palacios, G. *et al.* A new arenavirus in a cluster of fatal transplant-associated diseases. *N. Engl. J. Med.* **358**, 991–998 (2008).
82. Quan, P. L. *et al.* Astrovirus encephalitis in boy with X-linked agammaglobulinemia. *Emerging Infect. Dis.* **16**, 918–925 (2010).

83. Flygare, S. *et al.* Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biology* **17**, 111 (2016).
84. Naccache, S. N. *et al.* A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res.* **24**, 1180–1192 (2014).
85. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**, R46 (2014).
86. The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3807889/>. (Accessed: 10th December 2018)
87. In-Depth Investigation of Archival and Prospectively Collected Samples Reveals No Evidence for XMRV Infection in Prostate Cancer. Available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0044954>. (Accessed: 10th December 2018)
88. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* **12**, 87 (2014).
89. Brain Microbial Populations in HIV/AIDS:  $\alpha$ -Proteobacteria Predominate Independent of Host Immune Status. Available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0054673>. (Accessed: 10th December 2018)

90. Salzberg, S. L. *et al.* Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system. *Neurol Neuroimmunol Neuroinflamm* **3**, e251 (2016).
91. Ruby, J. G., Bellare, P. & Derisi, J. L. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda)* **3**, 865–880 (2013).
92. van Hal, S. J., Stark, D., Harkness, J. & Marriott, D. *Candida dubliniensis* Meningitis as Delayed Sequela of Treated *C. dubliniensis* Fungemia. *Emerg Infect Dis* **14**, 327–329 (2008).
93. Andrew, N. H., Ruberu, R. P. & Gabb, G. The first documented case of *Candida dubliniensis* leptomenigeal disease in an immunocompetent host. *BMJ Case Rep* **2011**, (2011).
94. Yamahiro, A., Lau, K. H. V., Peaper, D. R. & Villanueva, M. Meningitis Caused by *Candida dubliniensis* in a Patient with Cirrhosis: A Case Report and Review of the Literature. *Mycopathologia* **181**, 589–593 (2016).
95. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
96. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
97. A universal algorithm for sequential data compression - IEEE Journals & Magazine. Available at: <https://ieeexplore.ieee.org/document/1055714>. (Accessed: 10th December 2018)
98. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
99. Zhao, Y., Tang, H. & Ye, Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28**, 125–126 (2012).

100. Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies | Science Translational Medicine. Available at: <http://stm.sciencemag.org/content/6/224/224ra24>. (Accessed: 11th December 2018)
101. Pan, W., Gu, W., Nagpal, S., Gephart, M. H. & Quake, S. R. Brain Tumor Mutations Detected in Cerebral Spinal Fluid. *Clinical Chemistry* **61**, 514–522 (2015).
102. Vlamincx, I. D. *et al.* Circulating Cell-Free DNA Enables Noninvasive Diagnosis of Heart Transplant Rejection. *Science Translational Medicine* **6**, 241ra77-241ra77 (2014).
103. Fan, H. C. *et al.* Non-invasive prenatal measurement of the fetal genome. *Nature* **487**, 320–324 (2012).
104. Gu, W. *et al.* Noninvasive prenatal diagnosis in a fetus at risk for methylmalonic acidemia. *Genetics in Medicine* **16**, 564–567 (2014).
105. Vogelstein, B. & Kinzler, K. W. Digital PCR. *PNAS* **96**, 9236–9241 (1999).
106. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339**, 819–823 (2013).
107. Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
108. Hsu, P. D., Lander, E. S. & Zhang, F. Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell* **157**, 1262–1278 (2014).
109. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821 (2012).
110. Briese, T. *et al.* Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *mBio* **6**, e01491-15 (2015).

111. Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology* **29**, 908–914 (2011).
112. Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature Medicine* **20**, 548–554 (2014).
113. Zou, H. *et al.* Quantification of Methylated Markers with a Multiplex Methylation-Specific Technology. *Clinical Chemistry* **58**, 375–383 (2012).
114. Akhras, M. S. *et al.* Connector Inversion Probe Technology: A Powerful One-Primer Multiplex DNA Amplification System for Numerous Scientific Applications. *PLOS ONE* **2**, e915 (2007).
115. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O’Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.* **23**, 843–854 (2013).
116. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature Methods* **6**, 315–316 (2009).
117. Li, J. *et al.* Replacing PCR with COLD-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. *Nature Medicine* **14**, 579–584 (2008).
118. Didelot, A. *et al.* Competitive allele specific TaqMan PCR for KRAS, BRAF and EGFR mutation detection in clinical formalin fixed paraffin embedded samples. *Experimental and Molecular Pathology* **92**, 275–280 (2012).
119. Saiki, R. K. *et al.* Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354 (1985).

120. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *PNAS* **108**, 9530–9535 (2011).
121. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nature Methods* **10**, 623–629 (2013).
122. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
123. Oxnard, G. R. *et al.* Noninvasive Detection of Response and Resistance in EGFR-Mutant Lung Cancer Using Quantitative Next-Generation Genotyping of Cell-Free Plasma DNA. *Clin Cancer Res* **20**, 1698–1705 (2014).
124. Almoguera, C. *et al.* Most human carcinomas of the exocrine pancreas contain mutant c-K-ras genes. *Cell* **53**, 549–554 (1988).
125. Burner, G. C. & Loeb, L. A. Mutations in the KRAS2 oncogene during progressive stages of human colon carcinoma. *PNAS* **86**, 2403–2407 (1989).
126. Tam, I. Y. S. *et al.* Distinct Epidermal Growth Factor Receptor and KRAS Mutation Patterns in Non-Small Cell Lung Cancer Patients with Different Tobacco Exposure and Clinicopathologic Features. *Clin Cancer Res* **12**, 1647–1653 (2006).
127. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
128. Kleinstiver, B. P. *et al.* Broadening the targeting range of *Staphylococcus aureus* CRISPR-Cas9 by modifying PAM recognition. *Nat. Biotechnol.* **33**, 1293–1298 (2015).

129. Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759–771 (2015).
130. Granerod, J. *et al.* Challenge of the unknown. A systematic review of acute encephalitis in non-outbreak situations. *Neurology* **75**, 924–932 (2010).
131. Wang, D. *et al.* Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol.* **1**, E2 (2003).
132. Ran, F. A. *et al.* In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186–191 (2015).
133. Imperiale, T. F. *et al.* Multitarget stool DNA testing for colorectal-cancer screening. *N. Engl. J. Med.* **370**, 1287–1297 (2014).
134. Kinde, I. *et al.* TERT promoter mutations occur early in urothelial neoplasia and are biomarkers of early disease and disease recurrence in urine. *Cancer Res.* **73**, 7162–7167 (2013).
135. Li, M. *et al.* Sensitive digital quantification of DNA methylation in clinical samples. *Nat. Biotechnol.* **27**, 858–863 (2009).
136. Koh, W. *et al.* Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7361–7366 (2014).
137. Zheng, Z. *et al.* Anchored multiplex PCR for targeted next-generation sequencing. *Nat. Med.* **20**, 1479–1484 (2014).
138. Shin, H. *et al.* Variation in RNA-Seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. *PLoS ONE* **9**, e91041 (2014).

139. Lin, S., Staahl, B. T., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife* **3**, e04766 (2014).
140. Davanloo, P., Rosenberg, A. H., Dunn, J. J. & Studier, F. W. Cloning and expression of the gene for bacteriophage T7 RNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.* **81**, 2035–2039 (1984).
141. Zawadzki, V. & Gross, H. J. Rapid and simple purification of T7 RNA polymerase. *Nucleic Acids Res.* **19**, 1948 (1991).
142. Perez, F. & Granger, B. E. IPython: A System for Interactive Scientific Computing. *Computing in Science Engineering* **9**, 21–29 (2007).
143. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **9**, 90–95 (2007).
144. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).

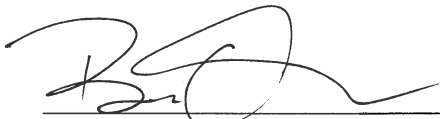


**Publishing Agreement**

*It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.*

***Please sign the following statement:***

*I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.*

  
\_\_\_\_\_  
Author Signature

12/21/2018  
Date