CLINICAL DIAGNOSTIC POTENTIAL AND CHARACTERIZATION OF DISTINCTLY HYPERMUTATED ANTIBODIES IN MULTIPLE SCLEROSIS PATIENTS

APPROVED BY SUPERVISORY COMMITTEE

Nancy L. Monson, Ph.D.

Lindsay G. Cowell, Ph.D.

Edward K. Wakeland, Ph.D.

Steven M. Patrie, Ph.D.

E. Sally Ward, Ph.D.

DEDICATION

I dedicate this thesis to my mother who instilled in me her academic curiosity and attention to detail that were so instrumental to the discovery of the many challenges introduced over the course of this project. I am also very grateful for the support I have received from my family and from my UTSW graduate class.

I would like to thank my mentor, Dr. Nancy Monson, for her continual support throughout my thesis work, as well as for her foresight in allowing me to expand the scope of my work and collaborations beyond the traditional boundaries of an Immunology Ph.D.

I would also like to thank my thesis chair, Dr. Lindsay Cowell, for her involvement in my training and her enthusiasm for bioinformatics discussions that helped shape the direction of my project.

Thank you to my thesis committee (Dr. Patrie, Dr. Ward and Dr. Wakeland) for their insight and advice over the course of my thesis.

CLINICAL DIAGNOSTIC POTENTIAL AND CHARACTERIZATION OF DISTINCTLY HYPERMUTATED ANTIBODIES IN MULTIPLE SCLEROSIS PATIENTS

by

WILLIAM HAROLD ALEXANDER ROUNDS

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

August 2016

CLINICAL DIAGNOSTIC POTENTIAL AND CHARACTERIZATION OF DISTINCTLY HYPERMUTATED ANTIBODIES IN MULTIPLE SCLEROSIS PATIENTS

WILLIAM HAROLD ALEXANDER ROUNDS, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2016

NANCY LEE MONSON, Ph.D.

Multiple sclerosis (MS) diagnosis primarily revolves around the use of brain lesion detection by MRI and the elimination of other possible neurological disorder diagnoses through clinical testing and history. For many patients first experiencing clinical symptoms that could be MS-related, this presents a challenge since diagnostic certainty based on clinical presentation and testing does not always reach a consensus among doctors who evaluate them.

With a growing body of evidence for B cell involvement and dysregulation in MS, our group investigated and identified a potential biomarker in the cerebrospinal fluid of patients with MS based on B cell antibody sequencing. This work first identified a distinct mutation pattern in the antibody sequences of CSF-derived B cells, termed the antibody gene signature (AGS), that could be used to identify patients with MS or patients who would convert to MS subsequent to their first onset of clinically detectable symptoms.

This thesis project outlines the transition from AGS testing in a laboratory setting to its use and implementation as an additional clinical diagnostic tool for MS (MSPrecise®) using next generation sequencing (NGS). One of its main goals is to thoroughly evaluate the performance of MSPrecise® using the far greater throughput which NGS allows for. Over the course of the project, NGS technology and accuracy optimization methods have advanced significantly. As our laboratory is the first to ever utilize NGS for somatic hypermutation evaluation, we focused strongly on the evaluation of challenges and features associated with NGS use for immune repertoire diversity and somatic hypermutation profiling of clinical samples. In this context, this project also highlights observations on sequence library preparation and post-sequencing data filtering that affect all immune repertoire research that uses these rapidly developing sequencing platforms.

TABLE OF CONTENTS

PUBLICATIONS

1. **Rounds W.H.**, Rivas J.R., Ireland S.J., Stowe A.M., Ligocki A.J., Guzman A.A., Ward E.S., Graves D., Frohman E.M., Greenberg B.M., Monson N.L. Antibodies with a unique mutation signature bind to autoantigens in relapsing remitting multiple sclerosis patients. **In preparation.**

2. **Rounds W.H.**, Wilks T.B., 2nd, Corboy J.R., Ratchford J.N., Murray R.S., Gudesblatt M., Bigwood D.W., Eastman E.M., Greenberg B.M., Cowell L.G., Monson N.L. Validation trial for a genetics-based add-on diagnostic test for multiple sclerosis. **In submission.**

3. Rivas J.R., Ireland S.J., **Rounds W.H.**, Chkheidze R., Lim J., Johnson J., Ramirez D., Ligocki A.J., Chen D., Guzman A.A., Wilson P., Meffre E., White C.L., 3rd, Greenberg B.M., Cowell L.G., Stowe A.M., Monson N.L. Peripheral VH4+ plasmablasts demonstrate autoreactive B cell expansion toward brain antigens in early multiple sclerosis patients. **In submission.**

4. **Rounds W.H.**, Salinas E.A., Wilks T.B., 2nd, Levin M.K., Ligocki A.J., Ionete C., Pardo C.A., Vernino S., Greenberg B.M., Bigwood D.W., Eastman E.M., Cowell L.G., Monson N.L. MSPrecise: A molecular diagnostic test for multiple sclerosis using next generation sequencing. Gene, 2015, November; 572(2): 191-7.

5. Ligocki A.J., Rivas J.R., **Rounds W.H.**, Guzman A.A., Li M., Spadaro M., Lahey L., Chen D., Henson P.M., Graves D., Greenberg B.M., Frohman E.M., Ward E.S., Robinson W., Meinl E., White C.L., 3rd, Stowe A.M., Monson N.L. A distinct class of antibodies may be an indicator of gray matter autoimmunity in early and established relapsing remitting multiple sclerosis patients. ASN Neuro, 2015, October; 7(5): 1759091415609613.

6. **Rounds W.H.**, Ligocki A.J., Levin M.K., Greenberg B.M., Bigwood D.W., Eastman E.M., Cowell L.G., Monson N.L. The antibody genetics of multiple sclerosis: comparing next-generation sequencing to Sanger sequencing. Frontiers in Neurology, 2014, September; 5:166.

7. Monson N.L., Ireland S.I., Ligocki A.J., Chen D., **Rounds W.H.**, Li M., Huebinger R.M., Cullum C.M., Greenberg B.M., Stowe A.M., Zhang R. Elevated CNS inflammation in patients with preclinical Alzheimer's disease. Journal of Cerebral Blood Flow and Metabolism, 2014, January; 34(1):30-3.

8. Ligocki A.J., **Rounds W.H.**, Cameron E.M., Harp C.T., Frohman E.M., Courtney A.M., Vernino S., Cowell L.G., Greenberg B., Monson N.L. Expansion of CD27high plasmablasts in transverse myelitis patients that utilize VH4 and JH6 genes and undergo extensive somatic hypermutation. Genes and Immunity, 2013, July; 14(5):291-301.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AA: amino acid

AGS: antibody gene signature

BBB: blood brain barrier

BCR: B cell receptor

CDMS: clinically definite multiple sclerosis

cDNA: complementary DNA

CDR: complementarity determining region

CIS: clinically isolated syndrome

CO: crossover (sample specific barcode contamination)

CNS: central nervous system

CSF: cerebrospinal fluid

DMT: disease modifying therapy

EAE: experimental autoimmune encephalomyelitis

EDSS: Expanded Disability Status Scale

FR: framework region

GC: germinal center

gDNA: genomic DNA

HC: healthy control

IFN: interferon

Ig: immunoglobulin

IgG: immunoglobulin of the gamma isotype

JH: junctional heavy segment

MBP: myelin basic protein

MF: mutation frequency

MOG: myelin oligodendrocyte glycoprotein

MRI: magnetic resonance imaging

MS: multiple sclerosis

NGS: next generation sequencing

NMO: neuromyelitis optica

OCB: oligoclonal band

OND: other neurological disorder

PB: peripheral blood

PCR: polymerase chain reaction

RMF: replacement mutation frequency

rhAb: full-length recombinant human antibody

RM: replacement mutation

RRMS: relapsing remitting multiple sclerosis

R:S ratio: replacement to silent mutation ratio

SHM: somatic hypermutation

SPMS: secondary progressive multiple sclerosis

VH: variable heavy segment

VH4: VH family 4

VLA-4: very late antigen-4

CHAPTER ONE

INTRODUCTION

Multiple Sclerosis (MS)

MS incidence and pathology

Multiple sclerosis (MS) is autoimmune inflammatory disease of the central nervous system (CNS) with an incidence of roughly 0.1% in the general population (Courtney et al., 2009). MS pathology is characterized by axonal damage from demyelination and the formation of lesions in the CNS (Trapp et al., 1998; Bitsch et al., 2000; Cepok et al., 2001; Geurts et al., 2005; Frohman et al., 2006a). Loss of the myelin sheath around neuronal axons leads to axonal damage and destruction, both at the site of lesions and in normal appearing white matter, causing permanent neurological damage (Trapp et al., 1998; Bitsch et al., 2000; Bjartmar et al., 2000; Bjartmar et al., 2001). MS symptoms are highly variable in function of which nerves are impacted by demyelination, which raises challenges for definitive clinical diagnosis.

Another diagnostic challenge is a growing body of evidence that damage in the CNS of MS patients can occur in both white matter and gray matter regions. The status of MS as solely a white matter disease is under investigation due to recent findings showing that white matter lesions are more readily detectable by magnetic resonance imaging (MRI) compared to gray matter lesions (Kidd et al., 1999; Bo et al., 2003). It has also been shown that gray matter atrophies at a greater rate than white matter in MS patients (Fisher et al., 2002; Chard et al., 2004; Valsasina et al., 2005; Fisniku et al., 2008a) and that the measurement of brain atrophy is a better predictor of clinical disease progression

(Sormani et al., 2013) than white matter damage alone (Moriarty et al., 1999; Bo et al., 2007). Recently, early-stage MS has been associated with gray matter demyelination (Lucchinetti et al., 2011) and binding to gray matter has been shown for a distinct subset of antibodies from MS cerebrospinal fluid (CSF) B cells (Ligocki et al., 2015).

Although it is unclear how MS is initiated, disruption of the blood brain barrier (BBB) has been identified as a key event that allows infiltration of lymphocytes into the CNS (Kirk et al., 2003; Leech et al., 2007; de Vries et al., 2012). MS lesions are characterized by infiltration of macrophages, T cells and B cells (Lucchinetti et al., 2000; Noseworthy et al., 2000; Frohman et al., 2006a; Lassmann et al., 2007) which can damage myelin by inducing a local inflammatory response. These cells are also observable in the CSF as frequencies of activated CD4 T cells and B cells in the CSF increase during active MS (Wang et al., 2002; Frohman et al., 2006a).

MS clinical presentation and subtypes

Initial presentation of MS occurs typically as an isolated episode of neurological disability which usually affects the optic nerves, brainstem or spinal cord (Miller et al., 2005). This demyelinating event is usually followed by phase of remission during which the patient recovers from the neurological symptoms (Lublin and Reingold, 1996; Courtney et al., 2009) (Figure 1-1). Patients experiencing these phases of acute attacks and remission are diagnosed with relapsing remitting multiple sclerosis (RRMS), which affects 85% of the total MS patient population. The other 15% experience a different disease course called primary progressive MS (PPMS), during which they experience a

progressive worsening of neurological symptoms without phases of recovery (Lublin and Reingold, 1996; Courtney et al., 2009).

After a median of 15 years, 66% of RRMS patients transition into a progressive form of the disease called secondary progressive MS (SPMS) and similar to PPMS with regards to a steady increase in disability with no remissions (Scalfari et al., 2010).

Current MS diagnostic toolbox

Patients that present with an initial episode of neurological deficits are diagnosed with clinically isolated syndrome (CIS) and are considered at varying risk to convert to clinically definite MS (CDMS), depending on a panel of criteria (McDonald et al., 2001; Polman et al., 2011). In order to be diagnosed with CDMS, patients must show signs of lesion dissemination in both time and space (McDonald et al., 2001; Polman et al., 2011; Milo and Miller, 2014). These current diagnostic criteria take advantage of the improvements in lesion type identification: T1 gadolinium enhancing lesions indicate active white matter lesions (Miller et al., 1988) and T2 gadolinium enhancing lesions are markers of prolonged disease activity (Molyneux et al., 1998). Although the current diagnostic criteria for RRMS have been improved, differentiating between RRMS and other neurological disorder (OND) remains a challenge and still relies on "the principle of no better explanation" (Milo and Miller, 2014) to rule out ONDs with similar clinical presentations.

In addition to radiological testing, detection of oligoclonal bands (OCB) in patient CSF was the first evidence for B cell dysregulation in MS (Kabat et al., 1942; Kabat et al., 1948; Kabat et al., 1950; Johnson and Nelson, 1977; Luxton et al., 1990). As a result,

since up to 90% of CIS and MS patients have OCBs in the CSF (Link and Muller, 1971; Jacobs et al., 1997; Freedman et al., 2005; Dobson et al., 2013) this test was incorporated in the early diagnostic criteria for CDMS (McDonald et al., 2001). However, it was also shown that OCBs can be found in OND patients (Link and Muller, 1971; Reske et al., 2005). Due to this limitation, the specificity of OCB has been revised at roughly 61% (Reske et al., 2005; Tintore et al., 2008; Petzold, 2013), which has limited its usage in subsequent updates to the diagnostic criteria (Polman et al., 2011; Milo and Miller, 2014).

The immediate consequence of the diagnostic complexity of MS is a prolonged time lapse before patients are able to start treatment (Milo and Miller, 2014). This is deleterious to long term prognosis for multiple reasons. Firstly, when CIS patients are initially identified, most show signs of ongoing damage in the CNS by MRI (Brex et al., 2002). Ongoing subclinical damage accumulation prior to a first attack is further supported by a recently discovered group of patients with similar MRI features as CDMS patients but who do not experience clinical symptoms, referred to as a radiologically isolated syndrome (RIS) (Moore and Okuda, 2009). These patients progressed to CIS in a median time of 5.4 years and highlights that disease progression is ongoing regardless of episodic clinical manifestations. The accumulation of an estimated 5-10 lesions per clinical relapse (Thrower, 2007) is further evidence of subclinical damage progression in CIS patients.

In this context, early diagnosis of MS is primarily beneficial in that it allows for earlier treatment. In fact, it has been shown that early treatment with disease modifying therapies (DMTs) delays both disease progression and accumulation of disability

(Frohman et al., 2006b; Fisniku et al., 2008b; Rocca et al., 2008; Scalfari et al., 2010; Greenberg, 2011; D'Alessandro et al., 2013). Delay of subsequent MS attacks has a profound effect on patient prognosis since increased rate of attacks in the initial years following a first clinical attack correlates with increases in disability (Confavreux et al., 2003; Scalfari et al., 2010; Gajofatto et al., 2013) and likelihood for SPMS conversion (Scalfari et al., 2010).

The role of B cells in MS

Antibodies in the CNS

The detection of OCB and elevated intrathecal immunoglobulin (Ig) in MS patient CSF was the first indication that antibodies had a role in MS (Kabat et al., 1942; Kabat et al., 1948; Kabat et al., 1950; Johnson and Nelson, 1977; Luxton et al., 1990). The heterogeneity of MS disease courses and clinical manifestations is echoed by the diversity of immune cell involvement in the lesions. MS lesion pathology analysis has identified four distinct patterns of immunological features and structure in active lesions (Lucchinetti et al., 2000). While patterns III and IV were characterized by oligodendrocyte loss and lack of remyelination, pattern I showed strong T cell infiltration and pattern II showed strong plasma cell infiltration as well as Ig and complement deposition. Patients with pattern II lesions respond well to plasmapheresis (Keegan et al., 2005; Magana et al., 2011), reinforcing the idea that antibodies in these patients have a pathogenic role. More specifically, Ig of the gamma isotype (IgG) co-localizes with complement C3b on demyelinated axons and oligodendrocytes and antibody-antigen complexes have been detected in lipid-loaded macrophages in active lesions (Sadaba et

al., 2012). Lastly, increases in intrathecal Ig (Sellebjerg et al., 2000; Izquierdo et al., 2002) and complement activation as measured by the terminal complement complex (Sellebjerg et al., 1998) correlate with neurological disability.

B cells in the CNS

As a result of the increase in BBB permeability in MS (Kirk et al., 2003; Leech et al., 2007; de Vries et al., 2012), the frequency of B cells is increased up to 17% in the CSF of MS patients (Cepok et al., 2005; Cepok et al., 2006; Ligocki et al., 2013) compared to less than 1% in healthy individuals (Svenningsson et al., 1995; Kleine and Benes, 2006; de Graaf et al., 2011). In normal human CNS, memory B cells express high levels of very late antigen-4 (VLA-4), a cell adhesion molecule, and are thus favored to cross the BBB compared to naïve cells (Kleine and Benes, 2006). In MS patient CSF, 80-85% of B cells have a memory phenotype ($CD19^{+}CD27^{+}$) (Cepok et al., 2006; Haas et al., 2011; Ligocki et al., 2013).

In addition to being increased in the CSF, B cells in the CNS of MS patients reside in structures called ectopic B cell follicles (Serafini et al., 2004; Magliozzi et al., 2007). These structures resemble lymphoid follicles and contain B cells, T cells, plasma cells and follicular dendritic cells, and express the lymphoid chemokines CXCL13 and CCL21 (Serafini et al., 2004) that are responsible for B cell trafficking to lymphoid tissue. Combined peripheral blood (PB) and CSF B cell repertoire analysis indicate B cell recruitment from the periphery into the CSF based on the identification of cross-compartment clonally related B cells (von Budingen et al., 2012). However, the presence

of ectopic B cell follicles in the CNS suggests that B cell maturation and selection can occur within this compartment.

Clinical data on existing DMTs point to multiple pathways of potential B cell involvement in MS. For example, Fingolimod is an antibody that prevents B and T cells from exiting secondary lymphoid tissue. MS patients treated with Fingolimod demonstrated decreased B cell counts in the periphery as expected, but treatment did not affect the CSF B cell population or intrathecal IgG (Kowarik et al., 2011). Nevertheless, disease progression as measured by relapse rate and new lesions by MRI was reduced (Kappos et al., 2010). In contrast, Natalizumab, an anti VLA-4 antibody was effective in reducing CNS B cell entry (Stuve et al., 2006; Kowarik et al., 2011) and reduced intrathecal IgG levels in the patients that remained relapse free (44%) (Villar et al., 2012). Lastly, treatment with Rituximab (Monson et al., 2005a; Cross et al., 2006; Hauser et al., 2008; Martin Mdel et al., 2009) or its humanized equivalent Ocrelizumab (Kappos et al., 2011) is effective in the treatment of RRMS and SPMS (Rommer et al., 2011) by depleting $CD20^+$ B cells in the CSF without a corresponding decrease in total intrathecal IgG (Cross et al., 2006; Piccio et al., 2010). This suggests that in patients who respond to Rituximab treatment, the primary involvement of CSF B cells is as antigen presenters.

Interestingly, Rituximab decreases autoantibody titers specifically, while leaving total IgG levels intact in other autoimmune diseases such as neuromyelitis optica (NMO) (Kim et al., 2011), vasculitis (Ferraro et al., 2008), systemic lupus erythematosus (Ioannou et al., 2008) and rheumatoid arthritis (Lazarus et al., 2012). The impact of autoantibody reduction by B cells on responder status in MS patients remains to be determined. However, in mouse models of MS, B cells producing antibodies that bind

myelin oligodendrocyte glycoprotein (MOG) drive residual disease following B cell depletion (Chen et al., 2014; Chen et al., 2016).

Distinct B cell receptor genetics in MS

B cell development and effector function is driven by cell surface antibody expression, i.e. the B cell receptor (BCR) (Meffre et al., 2000; Gauld et al., 2002). As a result, early efforts to understand the abnormal functions of B cells and their antibody products in MS CSF have focused on elucidating the genetics of the BCR. One of the most well characterized differences is the increase in frequency of variable heavy chain family 4 (VH4) gene family usage in RRMS patient CNS compared to expected distributions (Owens et al., 1998; Qin et al., 1998; Baranzini et al., 1999; Colombo et al., 2000; Owens et al., 2003; Monson et al., 2005a; Harp et al., 2007; Owens et al., 2007). This increase in VH4 is also observed in the OCBs (Baranzini et al., 1999).

BCR sequencing efforts also led to the discovery of extensive somatic hypermutation and clonal expansion in the CNS (Baranzini et al., 1999; Smith-Jensen et al., 2000; Owens et al., 2001) and CSF (Qin et al., 1998; Colombo et al., 2000; Colombo et al., 2003; Owens et al., 2003; Qin et al., 2003; Ritchie et al., 2004; Monson et al., 2005b; Harp et al., 2007; von Budingen et al., 2012) of MS patients. Some of these studies even identified clones that persisted over the course of one or more years (Colombo et al., 2000; Colombo et al., 2003).

In addition to clonal expansion, the hallmarks of germinal center (GC) selection are accumulation of mutation frequency, receptor editing, increased targeting of mutations to complementarity-determining regions (CDR) compared to framework

regions (FR) (Dorner et al., 1998), and increased ratio of replacement to silent mutations in the CDR (Owens et al., 1998). Previous work by our group has shown that MS CSF B cells undergo typical GC selection according to these criteria (Monson et al., 2005b; Harp et al., 2007). Since these VH4 enriched MS CSF B cells had all the genetic characteristics of response to antigen, we hypothesized that aberrant binding to autoantigen would correspond to a distinct pattern of somatic hypermutation. This study identified 6 codons that had significantly elevated replacement mutation frequency in CSF VH4 B cells compared to a healthy control PB B cell repertoire (Cameron et al., 2009). The combination of these 6 codon positions was dubbed the antibody gene signature (AGS) and was 91% accurate in identifying patients with RRMS or CIS patients who would convert to RRMS in this early study.

Collectively, the genetic features of MS CSF B cell antibody receptor sequences combined with this newly-discovered mutation-based biomarker for RRMS highlighted the potential usefulness of bulk repertoire analysis as a supportive tool for MS clinical diagnosis.

New tools for repertoire analysis in the clinic

Next-generation sequencing (NGS)

The advent of NGS technologies have reduced the cost sequencing and facilitated its broader diffusion (Shendure and Ji, 2008). These tools offer in depth repertoire characterization (Boyd et al., 2009; Boyd et al., 2010; Arnaout et al., 2011) and can be used to bypass single-cell sorting for Sanger sequencing when paired heavy and light chain sequence data is not required.

The majority of NGS full length VDJ recombinant segment sequence data presented in this thesis was generated using the Roche 454 NGS platform (Mardis, 2011). When this project was started, the 454 platform was the only reliable source of long sequence reads required for full length sequence coverage (with the addition of primers, reads longer than 350 base pairs are not uncommon) (Mardis, 2011). In contrast, the Illumina sequencing platform has recently improved sequence length and is now a viable alternative to 454 sequencing. Illumina's pros and cons compared to 454 will be outlined in the discussion section of this document. Both platforms share the main features of NGS compared to Sanger sequencing: they rely on the generation of a pooled library of DNA segments to be sequenced and then isolate these molecules from each other prior to performing hundreds of thousands of local sequencing reactions simultaneously, which is why these methods are described as "massively parallel sequencing".

Compared to Sanger, NGS outputs one sequence per final template, rather than a sequence that represents an average over many templates which masks mutations that occur in fewer than 20% of templates (Davidson et al., 2012). As a result, any polymerase chain reaction (PCR) and sequence-related errors are not averaged out, but rather are carried through all the way to end and are outputted in the final repertoire (Galan et al., 2010; Prabakaran et al., 2011). For this reason, expected sources of error must be carefully evaluated to optimize PCR protocols and data analysis filter criteria and tools as 454 sequencing has a substitution error rate about 10-fold higher than Sanger sequencing (Kircher and Kelso, 2010).

The primary distinguishing feature of NGS platforms is in the sequencing reaction and corresponding signal detection method. 454 sequencing relies on sequential cycles of

nucleotide addition, with each cycle allowing for more than a single nucleotide extension as long as all subsequent nucleotides match the first one. As a result, signal output from the 454 platform has to be evaluated at each cycle to determine the number of matching nucleotide incorporations (Figure 1-2). This platform has a lower nucleotide substitution rate than other comparable platforms, and instead is more likely to generate insertion and deletion (indel) errors, particularly in regions that contain stretches of 2 or more identical nucleotides (Bolotin et al., 2012), which it does at a reported frequency of 3.8-5 x 10^{-3} (Loman et al., 2012; Georgiou et al., 2014).

Antibody Gene Signature (AGS) and MSPrecise®

In order to be able to test and use the AGS in a clinical setting, transitioning from single-cell sorted CSF B cells to sequencing from bulk lymphocytes is critical to allowing any facility that can process blood to prepare a sample for AGS testing. Evaluating the impact of PCR and NGS error on somatic hypermutation identification is key to robust AGS score generation, which is entirely dependent on observed replacement mutations. Since previous work has focused on clonality monitoring rather than somatic hypermutation (SHM) evaluation (Boyd et al., 2009; Boyd et al., 2010; Arnaout et al., 2011; Logan et al., 2011), our first foray into NGS compared paired single-cell Sanger and NGS B cell antibody gene repertoires to evaluate the cross-platform robustness of the AGS (Rounds et al., 2014).

This preliminary evaluation identified key issues with the AGS protocol and data analysis pipeline. After correcting these (detailed in the methods), the AGS test was renamed MSPrecise® and re-assessed using a larger cohort of patients with long-term

RRMS (Rounds et al., 2015) and a very large validation cohort for better OND subgroup performance evaluation.

Challenges to autoantigen target identification in MS

Previous approaches to autoantigen identification

The identification of putative autoantigens in MS has been a driving goal of the field since such targets would provide a greater understanding of MS initiation and onset as well as facilitate diagnosis (summary Table 1-1). Early work on experimental autoimmune encephalomyelitis (EAE), the mouse model of MS, identified a myelin antigen MOG that could induce demyelination (Appel and Bornstein, 1964; Seil et al., 1968). Subsequent work showed that EAE could be transferred to another mouse through passive transfer of antibodies against MOG (Schluesener et al., 1987; Linington et al., 1988).

The use of EAE to model MS in mice and the importance of demyelination in MS progression has put most of the focus of autoantigen identification on MOG and other myelin proteins such as myelin basic protein (MBP) and proteolipid protein, all major components of the myelin in the CNS (Quarles, 2005). However, there have been conflicting reports in which anti-MOG antibodies correlate (Angelucci et al., 2005; Klawiter et al., 2010) or fail to correlate (Breij et al., 2006; Kuhle et al., 2007; Tewarie et al., 2012) with more severe markers of disease progression. Furthermore, anti-myelin antibodies are also found in OND patients (Karni et al., 1999) and healthy donors (Lampasona et al., 2004) and do not demonstrate high affinity binding to their target in MS (O'Connor et al., 2003).

The most common techniques for autoantigen identification use MS serum, CSF (neat or supernatant) or recombinant human antibodies (rhAbs) derived from single B cells (Fraussen et al., 2009). Often these screens rely on affinity proteomics approaches to identify specific targets among a large number of proteins, such as the one used to identify neuronal antigens as putative targets in MS. These include neurofascin (Mathey et al., 2007; Lindner et al., 2013), an axoglial protein against which antibodies were found in a subgroup of MS patients (Kawamura et al., 2013), and contactin-2 (Derfuss et al., 2009), another axoglial protein which did not display good sensitivity in a follow up study (7.8% in RRMS) (Boronat et al., 2012).

Another example of the challenges associated with autoantigen identification is the discovery in MS serum of anti-Kir4.1 antibodies in 47% of MS patients compared to 1% in OND and 0% in healthy donors (Srivastava et al., 2012). Unfortunately, subsequent work by two independent research teams failed to replicate these findings (Brickshawana et al., 2014; Nerrant et al., 2014). Another research team did find elevated levels of anti-Kir4.1 antibodies in MS patients and noted that anti-Kir4.1 serum titers were significantly higher during MS relapse than remission (Brill et al., 2015).

AGS-associated autoantigen screening

RRMS disease progression has been shown to correlate with detectable changes in the CSF such as high B cell frequencies associated with rapid progression (Cepok et al., 2001). In addition, RRMS patient CSF is a dynamic environment which has been shown to have fluctuating levels of putative autoantibodies (Brill et al., 2015). These compound the difficulty of identifying relevant autoantigen targets by screening the total

CSF antibody or B cell pool and help explain why definitive autoantigens for even subgroups of multiple sclerosis patients have yet to be validated.

In order to minimize some of these challenges and further characterize the AGS, we focused our efforts on the cloning and expression of clonally expanded and AGS-enriched MS CSF B cell antibody genes. Using a combination of molecular biology and proteomics methods, we screen a panel of rhAbs against a large number of human proteins with the goal of identifying promising targets for future validation.

Summary

The challenges associated with the clinical diagnosis of MS have fueled the scientific community's research into better understanding the immune response behind the disease, as well as its potential triggers. The aim of the work presented here is to further evaluate a potential BCR genetics biomarker identified in the CSF of RRMS patients that can identify patients with RRMS or who will convert to RRMS from those with an OND. In order to test this tool on a clinical scale, significant changes had to be made to the protocols first used to identify the AGS. These changes were required, both to eliminate the protocol's reliance on flow cytometry (a significant time and cost investment), and to optimize the use of NGS output for somatic hypermutation analysis and verification of the tools reliability using this new source of data. These changes were tested and implemented incrementally over the course of three distinct studies, with a driving goal to improve sequence data quality through changes to sequence library preparation and post-sequencing data filtering.

FIGURE LEGENDS FOR CHAPTER ONE: INTRODUCTION

Figure 1-1. Multiple sclerosis disease courses. X-axis represents disease progression over time. Y-axis represents relative levels of patient disability. Types of multiple sclerosis are indicated by the labelled arrows at the bottom of the graph with the start of the arrow indicating the first diagnosis position on the timeline that corresponds to that type of MS.

Figure 1-2. 454 signal processing. [Source: 454 Life Sciences, Roche] Signal intensity is evaluated at each nucleotide cycle to determine homopolymer length when applicable. Longer homopolymers have a greater likelihood of incorrect length evaluation, thus favoring insertion/deletion type errors over substitutions errors.

FIGURES FOR CHAPTER ONE: INTRODUCTION

Figure 1-1.

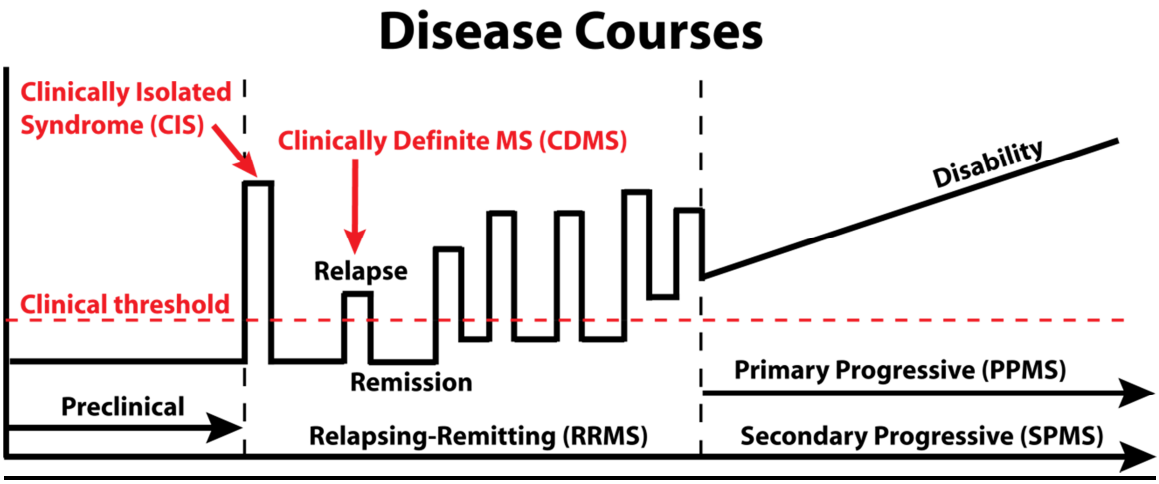


Figure 1-2.

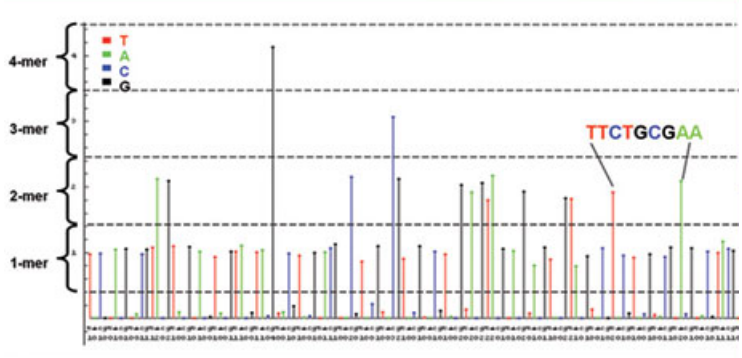


TABLE FOR CHAPTER ONE: INTRODUCTION

Table 1-1. Putative autoantigen targets in MS and conflicting reports

Antigen	Finding	References
Myelin antigens		
MOG	Can induce EAE	(Appel and Bornstein, 1964; Seil et al., 1968)
	Transfer of anti-MOG antibodies induces EAE	(Schluesener et al., 1987; Linington et al., 1988)
	Anti-MOG antibodies correlate with MS progression	(Angelucci et al., 2005; Klawiter et al., 2010)
	Anti-MOG antibodies don't correlate with MS progression	(Breij et al., 2006; Kuhle et al., 2007; Tewarie et al., 2012)
	Anti-MOG antibodies found in OND patients	(Karni et al., 1999)
	Anti-MOG antibodies found in healthy donors	(Lampasona et al., 2004)
MBP	MS anti-MBP antibodies are low affinity	(O'Connor et al., 2003)
Neuronal antigens		
Neurofascin	Anti-neurofascin antibodies induce axonal injury	(Mathey et al., 2007; Lindner et al., 2013)
	Anti-neurofascin antibodies found in subgroup of MS patients	(Kawamura et al., 2013)
Contactin-2	Anti-contactin-2 antibodies in MS patient serum (5/9)	(Derfuss et al., 2009)
	Anti-contactin-2 antibodies not in MS patient serum (4/51)	(Boronat et al., 2012)
Astrocyte antigen		
Kir4.1	Anti-KIR4.1 antibodies in MS patient serum (47%)	(Srivastava et al., 2012)
	Anti-KIR4.1 antibodies not in MS patient serum	(Brickshawana et al., 2014; Nerrant et al., 2014)
	Anti-KIR4.1 antibodies elevated in MS patient serum during relapse not remission	(Brill et al., 2015)

CHAPTER TWO

METHODS

Current data analysis pipeline

A driving force behind the work presented in this dissertation is the transition from single-cell Sanger methods of CSF B cell sequencing for mutation pattern scoring to a clinically practical CSF cell pellet that can shipped directly to a processing facility for DNA extraction, targeted amplification and sequencing. Initially, the primary difference between the two sequencing methods was in the interpretation of the output sequence data (Table 2-1). As a result of these differences, the methods used in the sequence data generation and analysis pipeline have undergone notable changes and revisions over the years and over multiple studies. These changes are outlined in Table 2-2 and the reasons for their implementation, as well as a more complete review of corresponding data analysis optimization will be presented in detail in Chapter 8 (Discussions). Overall, the transition from single-cell Sanger sequencing protocols, outlined in detail in a previously published doctoral dissertation (Ligocki, 2014), to the current NGS methods corresponds to a transition from sorted B cell complementary DNA-based (cDNA) multiple amplification PCR rounds to a single amplification PCR step using whole genome amplified DNA from a non-sorted CSF cell pellet (Table 2-2). The most up to date methods of the current data analysis pipeline are detailed below (and summarized in Figure 2-1), and as Table 2-2 shows are slightly modified from the methods used in the "Verification" study (Rounds et al., 2015).

NGS sequencing methods for all three studies (Chapter III, IV and V: "Confirmation", "Verification" and "Validation") are detailed below. Methods are organized by type and differences in which studies used different methods will be preceded by the corresponding chapter number in brackets **([III],[IV],[V],[All]**). Chapters III and IV methods have also been previously published (Rounds et al., 2014; Rounds et al., 2015). Primers for the Confirmation study are listed in Table 2-3. Primers for the Verification and Validation studies are proprietary to DioGenix and are not included here.

CSF sample preparation

[All] All CSF samples were collected by lumbar puncture in accordance with institutional review board-approved protocols at each site.

[III] Single CD19+ B cells were sorted into individual wells of a 96-well microtiter plate for single-cell Sanger DNA sequencing. At the same time, a pool of sorted CD19+ B cells from each patient was collected for NGS analysis.

[IV][V] Total CSF cell pellets were generated from 8-10 mL of freshly collected CSF by centrifugation at 400 x g and 4^{o}C within 1 hour of collection. The CSF supernatant was transferred to a fresh tube and frozen at -80^{o}C. The cell pellet was resuspended in 400 uL RPMI cell culture medium, transferred to a 2 mL cryovial and centrifuged again. The cell-free supernatant was discarded and the CSF cell pellet was frozen at -80^{o}C until use.

[All] Naïve (CD19+CD27-) and memory (CD19+CD27+) peripheral blood B cell pools were isolated from 3 healthy control samples and used as process controls to evaluate batch to batch variation and to aid in the evaluation of potential sequence errors generated during processing. Peripheral blood from healthy control donors was collected in blood tubes containing heparin as an anti-coagulant (BD, Franklin Lakes, NJ). Peripheral blood mononuclear cells (PBMCs) were isolated by centrifugation through Ficoll-Paque (GE Healthcare, PA). PBMCs were washed, counted and stained before being used to isolate naïve and memory B cells as described previously (Ireland et al., 2012). Total B cells were isolated from PBMC by magnetic activated cell separation (MACS) using a-CD19 microbeads (Miltenyi) according to the manufacturer's instructions. Purity was typically above 95%. Total CD19+ B cells were stained with CD19-PECy5, CD27-PE and IgD-FITC (BD Bioscience) and sorted into naïve (CD19+ IgD+ CD27-) and memory (CD19+ CD27+) populations on a FACS Aria (BD Biosciences, custom order system).

PCR and next generation sequencing of antibody genes from CSF cell pellets

[III] cDNA was amplified using a Primer Extension Preamplification (PEP) protocol as previously described by our group (Ligocki et al., 2013; Ligocki, 2014) and modified from an earlier protocol (Tiller et al., 2008). The detailed method for a full single-cell sorted plate with a 100 B cell well for NGS analysis is copied below from previously published methods (Ligocki, 2014). The 96-well plates were stored at -80°C post sort with cells frozen in 4 uL cPEP sort mix (10% 0.1M Dithiothreitol (DTT), 9% recombinant RNAsin (Promega) in 0.5x PBS). The cDNA was made within the same

plate, in a sterile RNA and DNA free environment including the storage space, working space, reusable and disposable components, equipment, and a dead-air hood. The plate was kept either on a bed of dry ice or on a metal plate holder previously stored at -20°C. 3.5 uL of the random hexamer primer (RHP) mix was added quickly to each well with a multi-channel pipette to prevent the mix from freezing in the pipette tip. This step also included a surfactant (10% Igepal CO-630) to break open the cell membrane without disrupting the nuclear membrane, resulting in clean access to mRNA. The plate was sealed and placed in a PCR cycler (Eppendorf Mastercyclers) and incubated for 1 minute at 68°C. After the incubation, the plate was removed and cooled on an ice block. The plate was then transferred to a new ice block and returned to the sterile RNA and DNA free environment. 7 uL of the reverse transcription (RT) mix was added to each well with a multi-channel pipette and mixed by pipetting 6-8 times. Each of the wells was then topped with 20 uL of mineral oil to protect and seal the reaction. The plate was sealed with a film and pulse-spun in a balanced mini-plate centrifuge or an Eppendorf centrifuge with plate adaptors. The cDNA reaction was run in the PCR cycler with a 42°C initial hot-start for 5 minutes followed by 25°C for 10 minutes, the annealing and extension cycle was at 50°C for 120 minutes and the reaction was completed and inactivated at 94°C for 5 minutes. Once completed, the plate with cDNA was stored at -20°C as template for downstream PCR amplifications. Table 2-4 has the formulations and component specifics for the cPEP sort, RHP, and RT mixes.

[IV] Genomic DNA (gDNA) was isolated using the QIAamp DNA Micro Kit (Qiagen, CA) and following the “Isolation of Genomic DNA from Small Volumes of Blood” protocol with a final elution volume of 15uL and quantitated using the Quant-iT

PicoGreen dsDNA Assay Kit (Invitrogen, Carlsbad, CA). Whole genome amplification (WGA) was performed using the REPLI-g Mini Kit (Qiagen, CA) protocol for "Amplification of Purified Genomic DNA using the REPLI-g Mini Kit" on up to 1000 cell equivalents of gDNA (6.6 ng) isolated from each clinical sample. Either the 2.5uL of gDNA or 5uL of gDNA version of this protocol was used depending on the gDNA concentration of each sample in order to meet the desired cell equivalent amounts. Each WGA reaction is cleaned using the QIAamp DNA Micro Kit (Qiagen, CA). DNA was quantitated using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Carlsbad, CA).

[V] Genomic DNA (gDNA) was isolated using the QIAamp DNA Micro Kit (Qiagen, CA) and following the "Isolation of Genomic DNA from Small Volumes of Blood" protocol with a final elution volume of 15uL. DNA was quantitated using the Qubit® dsDNA HS Assay Kit (ThermoFisher, MA) with 2uL of gDNA (if below detection threshold increase to 3uL). Whole genome amplification (WGA) was performed using the REPLI-g Mini Kit (Qiagen, CA) protocol for "Amplification of Purified Genomic DNA using the REPLI-g Mini Kit" on up to 2000 cell equivalents of gDNA (13.2 ng) isolated from each clinical sample. Either the 2.5uL of gDNA or 5uL of gDNA version of this protocol was used depending on the gDNA concentration of each sample in order to meet the desired cell equivalent amounts. DNA was quantitated using the Qubit® dsDNA HS Assay Kit (ThermoFisher, MA) with 2uL of gDNA. Each WGA reaction is cleaned using the QIAamp DNA Micro Kit (Qiagen, CA) with a maximum of 3ug per column. DNA was quantitated using the Qubit® dsDNA HS Assay Kit (ThermoFisher, MA) with 2uL of gDNA.

[All] PCR amplification of IGHV4 sequences was performed using the 4-primer Amplicon Tagging strategy developed by Fluidigm (South San Francisco, CA) to allow for multiplex sequencing. To allow for the incorporation of specific barcode sequences to the amplicons generated for each patient, the 5' ends of the forward internal primers were extended to include the common sequence 1 (CS1) tag (Fluidigm) and the 5' ends of the reverse internal primers were extended to include the common sequence 2 (CS2) tag (Fluidigm). Patient-specific barcode sequences were added to nested PCR amplicons by performing an additional PCR reaction using forward primers that contain the 454A primer, 4 nucleotide key, unique 10 nucleotide MID barcode, and CS1' sequences, and reverse primers that contain the 454B primer, 4 nucleotide key, unique 10 nucleotide MID barcode, and CS2' sequences. Patient-specific MID barcode sequences and the 454 primer sequences were then added in a single barcoding PCR reaction using purified PCR product. All VH4 and JH custom primers were synthesized by Integrated DNA Technologies (Coralville, IA). The 454/barcode primers were purchased from Fluidigm.

[III][IV] Four external and four internal PCR reactions were performed for each sample to increase the total amount of patient DNA processed and minimize the chance of any stochastic effects for CSF samples that have very low numbers of VH4-expressing B cells. All PCR reactions were performed using Phusion High-fidelity DNA Polymerase (New England Biolabs (NEB), Ipswich, MA).

[V] Three identical VH4-targeting PCR reactions were performed for each sample to increase the total amount of patient DNA processed and minimize the chance of seeing any stochastic effects for CSF samples that have very low numbers of VH4-expressing B

cells. All PCR reactions were performed using Q5 High-fidelity DNA polymerase (New England Biolabs (NEB), Ipswich, MA).

[III] Each external PCR reaction consisted of 3.0 uL of PEP cDNA, 10.0 uL 2X Phusion DNA Polymerase Master mix (NEB), 1.0 uL each of 10 µM pooled external forward and reverse PCR primers and water to bring the total volume to 20 uL. PCR cycling conditions were as follows: 98^{o}C for 3 minutes followed by 23 cycles of 98^{o}C for 10 seconds, 68^{o}C for 10 seconds, 72^{o}C for 10 seconds. The last 72^{o}C extension was extended to 10 minutes followed by a 4^{o}C hold. Each internal PCR reaction consisted of 3.0 uL DNA from the external PCR reaction, 10.0 uL 2X Phusion DNA Polymerase Master mix (NEB), 1.0 uL each of 10 µM pooled CS1/CS2-tagged internal forward and reverse PCR primers and water to bring the total volume to 20 uL. PCR cycling conditions were as follows: 98^{o}C for 1 minute followed by 10 cycles of 98^{o}C for 10 seconds, 68^{o}C for 10 seconds, 72^{o}C for 10 seconds then 21 cycles of 98^{o}C for 10 seconds, 72^{o}C for 10 seconds. The last 72^{o}C extension was extended to 10 minutes followed by a 4^{o}C hold.

[IV] Each external PCR reaction consisted of 125 ng of WGA DNA, 10.0 uL 2X Phusion DNA Polymerase Master mix (NEB), 1.0 uL each of 10 µM pooled external forward and reverse PCR primers and water to bring the total volume to 20 uL. PCR cycling conditions were as follows: 98^{o}C for 3 minutes followed by 23 cycles of 98^{o}C for 10 seconds, 68^{o}C for 10 seconds, 72^{o}C for 10 seconds. The last 72^{o}C extension was extended to 10 minutes followed by a 4^{o}C hold. Each internal PCR reaction consisted of 3.0 uL DNA from the external PCR reaction, 10.0 uL 2X Phusion DNA Polymerase

Master mix (NEB), 1.0 uL each of 10 µM pooled CS1/CS2-tagged internal forward and reverse PCR primers and water to bring the total volume to 20 uL. PCR cycling conditions were as follows: 98^{o}C for 1 minute followed by 10 cycles of 98^{o}C for 10 seconds, 68^{o}C for 10 seconds, 72^{o}C for 10 seconds; then 21 cycles of 98^{o}C for 10 seconds, 72^{o}C for 10 seconds. The last 72^{o}C extension was extended to 10 minutes followed by a 4^{o}C hold.

[V] Each VH4-targeting PCR reaction consisted of 250 ng of WGA DNA, 25 uL 2X Q5 DNA polymerase Master mix (NEB), final concentration of 0.5 µM forward and 0.5 µM reverse PCR primers (CS1/CS2-tagged) and water to bring the total volume to 50 uL. PCR cycling conditions were as follows: 98^{o}C for 30 seconds followed by 30 cycles of 98^{o}C for 10 seconds, 64^{o}C for 10 seconds, 72^{o}C for 10 seconds. The last 72^{o}C extension was extended to 2 minutes followed by a 4^{o}C hold.

[III][IV] 20ul of each of the PCR reactions for each subject was analyzed on a 2% agarose-TAE gel. PCR reactions that yielded a visible band of the appropriate size (320-350 bp) were gel purified using QIAquick Gel Extraction Kit (Qiagen, CA) and pooled for each subject. DNA was quantitated using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Carlsbad, CA). For each sample, 20 ng of purified amplicon was added to a single 50uL reaction containing specific MID-barcode primers and the appropriate buffers for Phusion High-fidelity DNA Polymerase (New England Biolabs (NEB), Ipswich, MA). PCR cycling conditions were as follows: 98^{o}C for 1 minute followed by 10 cycles of 98^{o}C for 10 seconds, 68^{o}C for 10 seconds, 72^{o}C for 10 seconds; then 21 cycles of 98^{o}C for 10 seconds, 72^{o}C for 10 seconds. The last 72^{o}C extension was

extended to 10 minutes followed by a 4^{o}C hold. 20ul of the PCR reaction for each subject was analyzed on a 2% agarose-TAE gel. PCR reactions that yielded a visible band of the appropriate size (450bp to 500bp) were gel purified using QIAquick Gel Extraction Kit (Qiagen, CA). DNA was quantitated using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Carlsbad, CA).

[V] PCR products were purified using the Agencourt AMPure XP kit (Beckman Coulter, IN). 50uL of amplicon from each PCR reaction was aliquoted into a 0.5mL microfuge tube. The Agencourt AMPure XP bottle was gently shaken to resuspend any settled magnetic particles. 90uL of AMPure XP reagent was added to each sample and mixed by pipetting until a homogenous mixture was formed. Samples were incubated at room temperature for 5 minutes and then placed on the Qiagen 12 tube magnet. After the sample cleared (approx. 1 minute) the solution was aspirated and discarded. Bead pellets underwent x2 30 second washes with 200uL 70% Ethanol. Off the magnet, 50uL of nuclease free water was added to each sample and mixed by pipetting up and down 10 times. Samples were returned to the magnet to separate the beads from the eluate. Eluate was then transferred to a new 0.5mL tube. 10uL of clean amplicon was added to a single 50uL reaction containing specific MID-barcode primers. CS1/CS2-tagged MID-barcode primers were ordered from the Fluidigm 454 Barcode Library and received as sample-specific primer pools on a plate. Prior to set up, the Fluidigm plate was thawed at room temperature, and centrifuged at 1000rpm for 60 seconds to remove any droplets from the seal. The Q5 mastermix was thawed at room temperature in the negative hood. Once thawed, the mastermix pulse vortexed for 10 seconds and spun down. For each sample, 25uL of mastermix and 7uL of nuclease free water were combined and aliquoted at 32uL

into PCR strip tubes. 8uL of each 454 MID-barcode primer pool was added to its designated PCR reaction. 10uL of clean target PCR amplicon was added to the tube containing the MID-barcode assigned to that individual sample. All reactions were mixed by pipetting up and down 5-10 times. PCR cycling conditions were as follows: 98^{o}C for 30 seconds followed by 10 cycles of 98^{o}C for 10 seconds, 68^{o}C for 10 seconds, 72^{o}C for 10 seconds; then 15 cycles of 98^{o}C for 10 seconds, 72^{o}C for 10 seconds. The last 72^{o}C extension was extended to 2 minutes followed by a 4^{o}C hold. For each sample, 5uL of barcoded amplicon was run on an E-gel (Invitrogen, CA) to determine the sample failure rate before proceeding to library preparation gels. Enough water and loading buffer were combined in a 2:1 ratio, respectively, to accommodate all the PCR amplicons to be run on the E-gels. 5µL of barcoded PCR amplicon was mixed with 15µL of the water / loading buffer mixture and mixed by pipetting up and down 5-10 times. Each sample was loaded, run, and imaged on the Invitrogen E-gel system according to the technical guide (version K December 12, 2008). Each sample with a visible band found in the 450bp to 500bp region of the E-gel, 20uL of the barcoded PCR amplicon was loaded on a 2% agarose-TAE gel. Bands were excised and gel purified using the QIAquick Gel Extraction Kit (Qiagen, Valencia, CA). DNA was quantitated using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Carlsbad, CA).

[All] Equimolar amounts of 454/barcode-tagged DNA from each sample were pooled and sequenced together. Prior to emulsion PCR (emPCR), the pooled DNA sample library was analyzed using the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) to confirm that the DNA fragment sizes in the pool were of the appropriate

length and that there was a minimal amount of short sequences, e.g. primers and primer dimer. The pooled DNA was then used for emPCR and sequenced on the 454 GS FLX DNA Sequencer at SeqWright Genomic Services (Houston, TX) using the 454 Titanium chemistry (Roche/454, Branford, CT) according to the manufacturer's recommended protocols.

NGS 454 data processing

[III] Each unique sequence was aligned to germline gene segment sequences using the IMGT/HighV-QUEST tool (Alamyar et al., 2012). IMGT outputs were compiled using a Perl program developed at UTSWMC (Ligocki et al., 2010; Ligocki et al., 2013). All subsequent data processing steps described here were performed using a combination of Perl and SQL database programs developed in-house. Initial filtering removed any sequence which met at least one of the following criteria: out of frame, truncated read length, less than 85% homology to germline sequence, and alignment errors (as indicated by IMGT). Because the NGS sequences were generated from pools of 100 or fewer B cells per sample, we found certain sequences to be highly amplified. Combined with processing error rates (both PCR and 454 platform based), this results in some sequences that are found in multiple samples, which we termed sequence crossover.

Since CDR3 sequences of the *VH* chains should not match from sample to sample (Jackson et al., 2013), we adopted a strategy to identify sequences in multiple samples that share the same CDR3 subsequence. In order to properly filter these out, we needed to remove crossover sequences with exact CDR3 sequence matches, but also include highly similar CDR3s which fall into 3 categories that are still in-frame: up to 3 single

mismatches, a homopolymer insertion plus deletion, and 3 homopolymer insertions or deletions. By using Levenshtein distance comparison that measures the number of indels and mismatched nucleotides that separate each CDR3 nucleotide sequence pair (Levenshtein, 1966; Pieterse and Black, eds. 22 August 2013; Available from: http://www.nist.gov/dads/HTML/Levenshtein.html), we were able to cluster the crossover CDR3s using a maximal distance of 3, to account for the rare in-frame homopolymer indels. This matched previous work done on T cell receptor CDR3 sequences (Bolotin et al., 2012). Any sequence cluster present in multiple samples was removed from all sample sequence pools. The exceptions were clusters with ≥ 99% representation in a single sample, in which case we used this conservative cut-off to justify the sample source of the crossover sequence cluster, and only removed its members from the other sample databases.

[IV][V] Raw sequences and their corresponding quality information were uploaded to the VDJServer online repertoire analysis tool (https://vdjserver.org/). Using the VDJpipe tool, reads were trimmed of barcode sequences, filtered for a minimum mean sequence quality of 35 and for a minimum length of 200 nucleotides, and then aligned to each other within a sample to identify identical reads and collapsed into a single read when exactly matching. The number of copies of each identical sequence was kept as a sequence tag for subsequent analysis and filtering. Each sequence was then aligned to germline gene segment sequences using the IgBlast aligner (Ye et al., 2013) through VDJServer.

VDJServer sequence alignment data is output as a single tab-delimited file per sample, with one row used per sequence. To streamline data analysis of millions of reads,

a set of scripts was written to transfer this data into a format designed for the query language SQL. SQLite was used since it is open-source and file-based rather than server based. In SQLite, sequence features that are identified by alignment are directly connected to the sequence they describe, thus allowing for easy database filtering and querying based on combinations of features.

Initial filtering removed any sequence which met at least one of the following criteria: frame-shifting insertions or deletions, out of frame junction, stop codon present, truncated read length, less than 85% homology to germline sequence, missing CDR3, missing read coverage between Chothia-numbered codons 31-92 (Chothia and Lesk, 1987; Al-Lazikani et al., 1997), not VH4 aligned. We also removed highly amplified sequences that were present in multiple samples (identified by their matching CDR3 nucleotide segment) because CDR3 segment matching from sample to sample should not occur (Jackson et al., 2013). The exceptions were matching CDR3 containing sequences with $\geq$ 99% representation in a single sample, in which case we used this conservative cut-off to justify the sample source of the sequence, and only removed its CDR3 matches from the other sample databases.

[IV] We discarded unique sequence reads which had fewer than two copies in the raw sequence data.

[V] Although both forward and reverse sequencing data was generated for the studies included in this thesis, reverse reads were used for all data analysis as these reads had better quality coverage of the CDR3 region, which was used as key feature for data analysis. A Perl script that converts VDJpipe matching sequence count output and

IgBLAST repertoire characterization through VDJServer (May 2016 version) into an SQLite table has been included in Appendix 1. This script was designed specifically for use with the Chapter 5 study and features a Kabat region conversion tool for FR3 to limit the mutation information to codon 92 or below (V gene alignment normally extends into the CDR3 by several codons). It also takes advantage of the parallelization opportunities provided by the combination of 6 separate NGS runs that were required to generate 182 samples worth of sequencing data. The 6 SQLite tables were then combined into a single "SequenceRAW" table and run through a "cleaning" SQLite script (Appendix 2) that was written to remove sequences that don't fit specific filter criteria outlined below. All sequence filtering steps are also outlined in Figure 2-2, with the detailed breakdown of filtered sequence counts at each step.

Post-alignment, unique reads were defined as a combination of specific variable heavy segment (VH) gene, junctional heavy segment (JH) gene, CDR3 nucleotides and replacement mutations, rather than requiring a perfect nucleotide match for sequences to be collapsed into one. When combined, reads identified as belonging to one unique template had their copy count numbers added to track the exact correspondence between a unique read and the number of raw reads that were collapsed into it. Lastly, we discarded unique sequence reads which had fewer than two copies in the raw sequence data.

[IV] Samples were filtered prior to analysis: we required at least 10 unique reads after filtering for a sample to be included in our analysis (Table 1) and also required each samples to have at least >=55% of the raw reads be VH4 and >=50% of the raw reads not

be removed due to matching CDR3s from another sample (these two filters were passed by all samples with at least 10 unique reads).

[V] Samples were filtered prior to analysis: repertoires with 8 or fewer reads were labelled as low sequence. We also found that samples with high amplification bias (measured by sequence coverage ratio of the 2nd highest / the highest single sequence equal to less than 2%, termed "Coverage Ratio") had greater levels of background unique sequences likely derived from a single amplified template. As a result, samples with CR < 2% with 24 or fewer reads were also labelled as low sequence. Also, samples with >50% of the raw reads removed due to matching CDR3s from another sample were excluded from the study. This threshold was picked because of a clear separation of percentages across that cut-off: the lowest % above it was nearly twice as high as the highest % below it. Sample filtering and subsequent repertoire analysis was performed using an "analyze" SQLite script (Appendix 3) that was written to remove samples that didn't fit the filter criteria outlined above and to output the sample-level genetics information used in the study.

Mutation analyses

[All] Unique VH4 sequences were analyzed in the region between codons 31 and 92 following the Chothia numbering system (Chothia and Lesk, 1987; Al-Lazikani et al., 1997) using the framework (FR) and complementarity determining regions (CDR) originally defined by Kabat (Kabat et al., 1992). Mutation analyses were performed both at the nucleotide level and codon level. Mutations in a codon that resulted in an amino acid substitution are referred to as replacement mutations (RM) and mutations in a codon

that don't cause an amino acid substitution are called silent mutations (SM). The replacement mutation frequency (RMF) at each codon is the basis for calculating antibody gene signature (AGS) and MSPrecise scores: AGS uses codons 31b; 40; 56; 57; 81 and 89 whereas the updated MSPrecise tool uses codons 31b; 40; 56; 57 and 81. The scores are calculated as a sum of [RMF at the codon minus the average RMF (1.6) in a healthy control peripheral blood database] divided by the standard deviation (0.9) of the average RMF of the same healthy control database (Cameron et al., 2009). Patients with scores above 5.8 are identified as "RRMS".

Statistical analyses

Statistical analyses were done using GraphPad Software 6.00 (San Diego, California, USA, www.graphpad.com). Mutation frequencies and MSPrecise® scores were compared across cohorts by Mann Whitney test (statistical significance for all methods was attributed to p-values ≤ 0.05). VH4 and JH gene segment distributions were compared between cohorts using a chi-squared test of independence. Percentage deviation of individual genes relative to a uniform distribution were calculated as (observed - expected) / expected. Specificity, sensitivity and accuracy were calculated for MSPrecise® based on adjudicated diagnosis. Specificity was calculated as (# correct OND assessments) / (# OND samples); sensitivity was calculated as (# correct RRMS assessments) / (# RRMS samples); and accuracy was calculated as (# correct assessments) / (# samples). The diversity index (DI) of each sample was calculated as the Shannon entropy of its VH4 gene frequency distribution using the 8 common VH4 genes (VH4-30, VH4-31, VH4-34, VH4-38, VH4-39, VH4-4, VH4-59, VH4-61; excluding

VH4-28 [0.09%] and VH4-55 [0.05%] for near 0 alignment frequency and excluding alignments to open reading frame OR15-8 [0.06%]).

Illumina data analysis pipeline

B cell receptor sequencing is performed by Illumina sequencing of pooled B cells (Figure 2-3). First, whole genome amplification is performed directly on cell pellets of 1000 B cells by REPLI-g mini kit (QIAGEN). PCR amplification is performed using the widely used BIOMED2 primer system (van Dongen et al., 2003) and the high-fidelity AmpliTaq Gold polymerase (Applied Biosystems). Specifically, 100ng of the amplified REPLI-g DNA for a sample is used for each of four PCR reactions targeting VH gene families at the FR1 region (VH1 alone, VH3 alone, VH4 alone, and VH2/VH5/VH6 combined; FR1 primer for VH7 is redundant and not required) and each using the BIOMED2 JHconsensus primer. This targeted amplification is run for 35 cycles at 60C annealing temperature in 50uLs. Gel purification of each whole reaction volume is done using a 2% agarose gel and a final elution volume of 20uL.

Sequencing is done through HudsonAlpha (Huntsville, AL) using the MiSeq Illumina platform at 250bp paired-end reads. To enable this, sequencing adapters are added to the final PCR products by a 10 cycle reaction that uses combined primers: i.e. the FR1 BIOMED2 primers are modified on the 5' end by the addition of VH and JH adapters (a set of proprietary adapters developed between Dr. Monson and HudsonAlpha). This adapter adding PCR reaction uses 5uL of PCR product (25% of the total). Gel purification is repeated as before with a final elution volume of 20uL. 5uL of each of the four gel purified products are combined into a final 20uL sample that has the full VH family repertoire. HudsonAlpha performs quality control (QC) on the sample for both concentration by Qubit and sample integrity by 2100 Bioanalyzer, with the addition

of a Kapa RT-PCR on a 10nM dilution to validated concentration for sequencing. HudsonAlpha also performs the standard Illumina chastity filter prior to generating a final fastq output, which we when then subject to VDJserver analysis and sequence filtering detailed below.

Fastq sequence paired-end reads are merged and filtered out below a minimum merged length of 200 base pairs and a minimum merged quality of 40. Reads without proper recognition and subsequent trimming of primer sequences are also discarded. Finally, reads with matching CDR3s and VH JH gene alignment across multiple samples were removed as likely due to background error (affects <1% of total reads).

FIGURE LEGENDS FOR CHAPTER TWO: METHODS

Figure 2-1. 454 data analysis pipeline. Steps from DNA template source to the final step of 454 sequencing are represented. Box color represents the type of steps performed and delineates the transitions between the different teams involved in the project.

Figure 2-2. 454 sequencing data filtering summary. Filtering steps and corresponding filtered sequence numbers using the current sequence filtering pipeline (Chapter 5, "Validation" study; Appendixes 1-3). % raw corresponds to the total raw sequences remaining from the initial 454 raw sequence data output (a sequence with 10-fold sequence coverage, i.e. redundancy 9, is tracked as 10 raw sequence counts). % unique corresponds to the total of unique sequences, initially defined as exact nucleotide matches and subsequently redefined as "matching VHgene, JHgene, CDR3nucleotides, RM position from amino acid X to amino acid Y".

Figure 2-3. Illumina data analysis pipeline. Steps from DNA template source to the final step of Illumina sequencing are represented. Box color represents the type of steps performed and delineates the transitions between the different teams involved in the project.

FIGURES FOR CHAPTER TWO: METHODS

Figure 2-1.

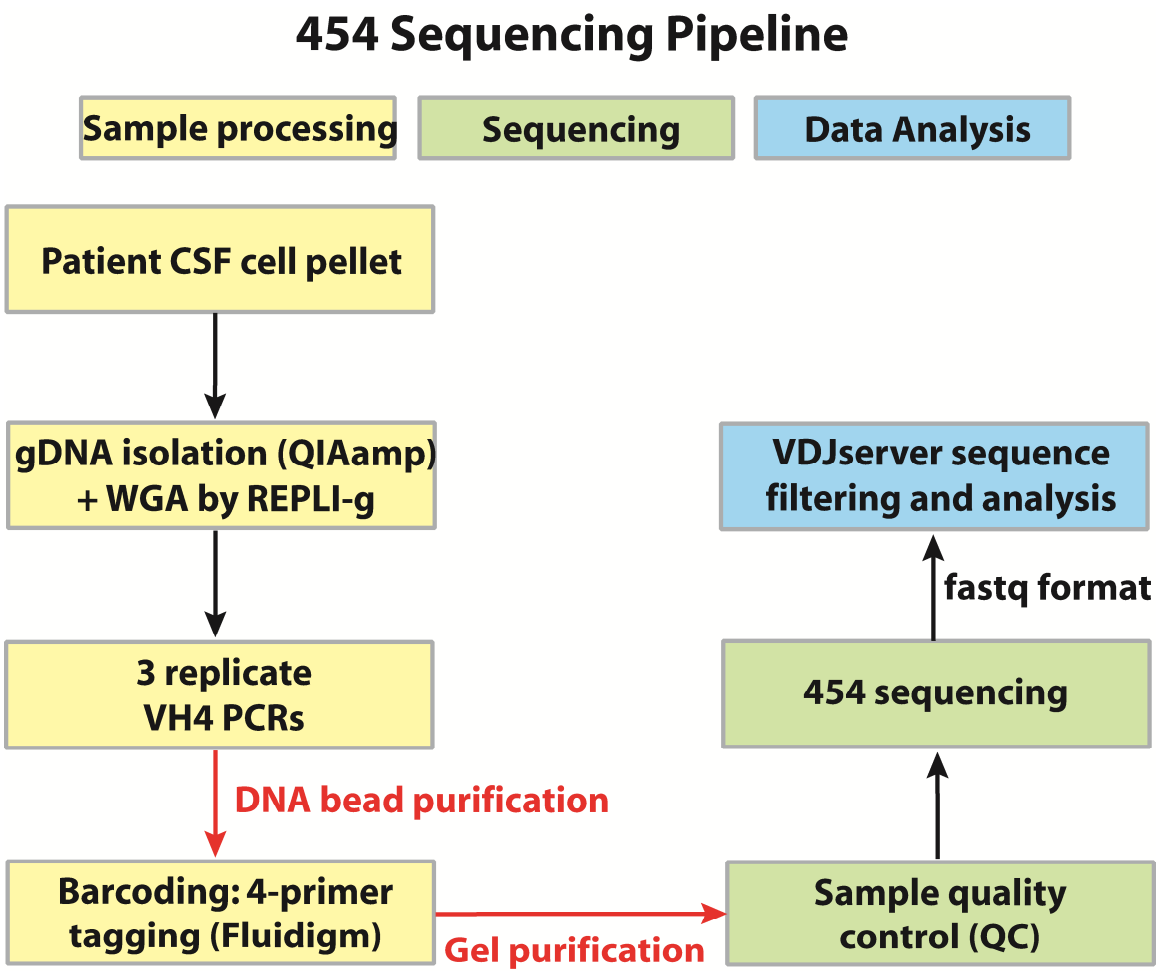
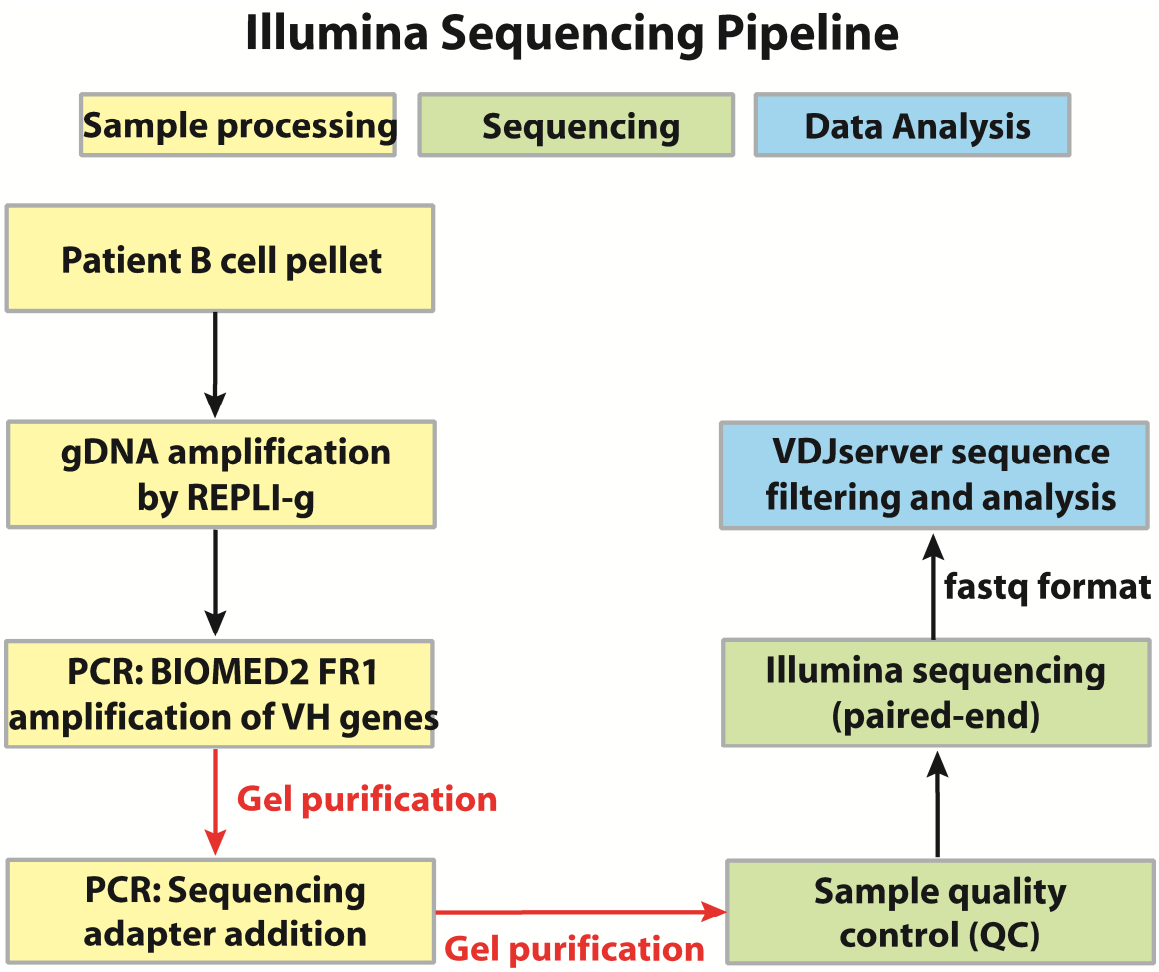


Figure 2-2.

	Validation Study Filtering Summary	counts	% raw	% unique
	Samples with any sequences	182		
	Raw sequences	1,854,595	100.0%	
VDJpipe Filter →	**Sequence quality filter (min length 200bp, min average quality window 35)**			
	Raw sequences that pass quality filters	1,657,349	89.4%	
	Unique sequences (exact nucleotide match)	307,583		100.0%
	↓ **IgBLAST**			
Sequence Filter 1 →	**REMOVE - Filter criteria (more than 1 per sequence possible)**			
	Not VH4 (raw)	31,550		
	Non VH alignments (raw)	750		
	N nucleotides in the sequence (raw)	1,022		
	No CDR3 reported (raw)	89,999		
	Missing Cystein or Tryptophan at CDR3 junction (raw)	185,573		
	CDR3 junction out of frame (raw)	174,833		
	Stop codons (raw)	233,541		
	insertions/deletions (raw)	230,005		
	Filtered sequences Stage 1			
	Raw sequences	1,174,124	63.3%	
	Unique sequences (exact nucleotide match)	153,809		50.0%
Sequence Filter 2 →	**REMOVE - Filter criteria 2 (applied sequentially)**			
	Truncated CDR or FR regions (raw)	3,365		
	Homology < 85% (raw)	23,911		
	Filtered sequences Stage 2			
	Raw sequences	1,146,848	61.8%	
	Matching sequences	151,488		49.3%
	↓ **New unique sequence definition**			
	Unique sequences (matching VHgene, JHgene, CDR3nucleotides, RM position from amino acid X to amino acid Y)			
	Unique sequences	49,074		100.0%
Crossover Filter →	**REMOVE - Crossover CDR3s filter (matching CDR3 nucleotides between samples)**			
	Crossover (raw)	143,188		
	Crossover (unique)	8,050		
	Unique CO removed sequences			
	Raw sequences	1,003,660	54.1%	
	Unique sequences	41,024		83.6%
Redundancy Filter →	**REMOVE - Redundancy 0 sequences (only 1 matching unique read found)**			
	R0 sequences (raw/unique)	24,859		
	↓ **Final sequence counts**			
	Final Unique sequences (Filtered, CO removed and Redundancy 1+)			
	Raw sequences	978,801	52.8%	
	Unique sequences	16,165		32.9%

Figure 2-3.


Illumina Sequencing Pipeline
Sample processing
Sequencing
Data Analysis
Patient B cell pellet
gDNA amplification by REPLI-g
PCR: BIOMED2 FR1 amplification of VH genes
Gel purification
PCR: Sequencing adapter addition
Gel purification
Sample quality control (QC)
Illumina sequencing (paired-end)
fastq format
VDJserver sequence filtering and analysis

TABLES FOR CHAPTER TWO: METHODS

Table 2-1. Data output differences between Sanger and 454 sequencing

Sanger sequencing	454 sequencing
1 read = averaged templates	1 read = 1 template
Nucleotides read one at a time	Homopolymers read together
1 starting template = 1 final read	1 starting template = 1000s of reads

Table 2-2. Primary process changes across three NGS AGS studies

Study	Chapter III[1]	Chapter IV[2]	Chapter V
Alternate name	"Confirmation"	"Verification"	"Validation"
Diagnosis	reviewed by one adjudicator	reviewed by one adjudicator	3 independent adjudicators
Source	CD19 CSF B cell pellet	CSF cell pellet	CSF cell pellet
DNA type	cDNA	gDNA	gDNA
PCR steps	External PCR	External PCR	
	Nested PCR	Nested PCR	Nested PCR
	Barcode PCR	Barcode PCR	Barcode PCR
Alignment tool	IMGT HighV-Quest	IgBlast through VDJServer	IgBlast through VDJServer
Unique read definition	Primer trimmed, exact nucleotide match	Primer trimmed 5 edge nucleotide mismatch	VH gene + JH gene + CDR3 nucleotides + RM pattern
"Unique" sequence coverage filter	Any	Minimum 2 reads per "unique"	Minimum 2 reads per "unique"
CDR3 crossover reads	Levenshtein distance max of 3 on AA: <99% in one sample = removed	Nucleotide exact match: <99% in one sample = removed	Nucleotide exact match: <99% in one sample = removed
AGS / MSPrecise	6 codon score	5 codon score	5 codon score

[1] (Rounds et al., 2014)
[2] (Rounds et al., 2015)

Definitions:

Primer trimming used in Chapter IV created sequences with a few additional edge nucleotides due to primer slipping. As a result, to prevent read matching issues to lower sequence coverage values, read matching was performed allowing for up to 5 combined nucleotide mismatches at either end.

RM pattern used in Chapter V meant that reads were not required to be exact nucleotide matches to be collapsed into one. Instead, reads with matching RMs from Chothia codons 31-92 were considered to be matching as long as VH gene, JH gene and CDR3 nucleotides also matched.

Table 2-3. Confirmation study PCR primer sequences

Primer name	Primer sequence	Direction
Primers for Confirmation study (Chapter III)		
External primers		
VH4E_1	caggagtggggcccag	5' Primer (5' → 3')
VH4E_2	caggagtcgggcccag	5' Primer (5' → 3')
VH4E_3	cagcagtggggcccag	5' Primer (5' → 3')
VH4E_4	cagcagtcgggcccag	5' Primer (5' → 3')
VH4E_5	caggagtggggcgcag	5' Primer (5' → 3')
VH4E_6	caggagtcgggcgcag	5' Primer (5' → 3')
VH4E_7	cagcagtggggcgcag	5' Primer (5' → 3')
VH4E_8	cagcagtcgggcgcag	5' Primer (5' → 3')
VH4E_9	caggagtggggctcag	5' Primer (5' → 3')
VH4E_10	caggagtcgggctcag	5' Primer (5' → 3')
VH4E_11	cagcagtggggctcag	5' Primer (5' → 3')
VH4E_12	cagcagtcgggctcag	5' Primer (5' → 3')
JHE_1	ctgaagagacagtgac	3' Primer (5' → 3')
JHE_2	ctgaagagacggtgac	3' Primer (5' → 3')
JHE_3	ctgaggagacagtgac	3' Primer (5' → 3')
JHE_4	ctgaggagacggtgac	3' Primer (5' → 3')
Nested primers		
VH4N_1	CS1-ggcccaggactggtgaagcctt	5' Primer (5' → 3')
VH4N_2	CS1-ggcgcaggactggtgaagcctt	5' Primer (5' → 3')
VH4N_3	CS1-ggctcaggactggtgaagcctt	5' Primer (5' → 3')
VH4N_4	CS1-ggcccaggactgttgaagcctt	5' Primer (5' → 3')
VH4N_5	CS1-ggcgcaggactgttgaagcctt	5' Primer (5' → 3')
VH4N_6	CS1-ggctcaggactgttgaagcctt	5' Primer (5' → 3')
JH1245N_1	CS2-gtgaccatggtcccttggccc	3' Primer (5' → 3')
JH1245N_2	CS2-gtgaccattgtcccttggccc	3' Primer (5' → 3')
JH1245N_3	CS2-gtgaccgtggtcccttggccc	3' Primer (5' → 3')
JH1245N_4	CS2-gtgaccgttgtcccttggccc	3' Primer (5' → 3')
JH36N_1	CS2-tgaccagggtgccacggccc	3' Primer (5' → 3')
JH36N_2	CS2-tgaccagggtgccccggccc	3' Primer (5' → 3')
JH36N_3	CS2-tgaccagggttccacggccc	3' Primer (5' → 3')
JH36N_4	CS2-tgaccagggttccccggccc	3' Primer (5' → 3')
JH36N_1	CS2-tgaccagggtgccatggccc	3' Primer (5' → 3')
JH36N_2	CS2-tgaccagggtgccctggccc	3' Primer (5' → 3')
JH36N_3	CS2-tgaccagggttccatggccc	3' Primer (5' → 3')
JH36N_4	CS2-tgaccagggttccctggccc	3' Primer (5' → 3')
CS1 = Fluidigm common sequence 1 tag CS2 = Fluidigm common sequence 2 tag		

Table 2-4. Formulations for mixes used for one cPEP plate (Ligocki, 2014)

Table 2-4. Formulations for mixes used for one cPEP plate[1]			
cPEP sort mix			
Component	*Vendor*	*Volume*	
0.1 M DTT (from SSRT III kit)[2]	Invitrogen/ Life-Technologies	53.7 uL	
Recombinant RNAsin (40U/uL)	Promega	48.3 uL	
0.5x PBS	Cellgro	454 uL	
		536.3 uL	***Final volume[3]***
		4 uL	***volume per well***
Random Hexamer Primer (RHP) mix			
Component	*Vendor*	*Volume*	
Random Hexamer primers (3 ug/uL)	Invitrogen/ Life-Technologies	7.67 uL	
10% Igepal CO-630	Sigma	76.7 uL	
Recombinant RNAsin (40U/uL)	Promega	23.33 uL	
Nuclease-free water	Biotex	392 uL	
		499.7 uL	***Final volume[3]***
		3.5 uL	***volume per well***
Reverse Transcription (RT) mix			
Component	*Vendor*	*Volume*	
5x First strand buffer (from SSRT III kit)[2]	Invitrogen/ Life-Technologies	396 uL	
10 mM dNTP	Promega	261.6 uL	
Recombinant RNAsin (40U/uL)	Promega	26.4 uL	
Nuclease-free water	Biotex	8.4 uL	
Superscript RT III (200U/uL)	Invitrogen/ Life-Technologies	33 uL	
		857.4 uL	***Final volume[3]***
		7 uL	***volume per well***

[1] One cPEP plate (96 wells) contained both the sorted cells and the cPEP reaction, which was performed in the same sorted plate.
[2] These reagents were provided as part of the kit with the Superscript RT III (SS RT III) enzyme.
[3] Final volume of the mix accounted for volume loss due to pipetting and transferring to reagent reservoirs.

CHAPTER THREE

RESULTS

AIM I: AGS scoring by next-generation sequencing is a reliable replacement for MS conversion diagnosis by single-cell Sanger sequencing analysis

Overview and rationale

Single-cell Sanger sequencing limits the clinical testing of AGS due to its cost, reliance on flow cytometry and slow turnaround time. In order to provide AGS evaluations to patients as part of their diagnostic workup, we developed protocols to generate AGS scores using next-generation sequencing (NGS) on CSF-derived cell pellets without the need to isolate single cells. No investigations have focused on whether NGS-based repertoires will properly reflect antibody gene frequencies and somatic hypermutation patterns defined by Sanger sequencing. Thus, the goal of this study was to evaluate whether NGS could adequately reflect Sanger SHM detection. CSF samples with paired single-cell sorted and pooled CD19+ cell pellets were used for this analysis.

THE ANTIBODY GENETICS OF MULTIPLE SCLEROSIS: COMPARING NEXT-GENERATION SEQUENCING TO SANGER SEQUENCING.

The following study has been published in *Frontiers in Neurology*. Rounds WH, Ligocki AJ, Levin MK, Greenberg BM, Bigwood DW, Eastman EM, Cowell LG, Monson NL, *The antibody genetics of multiple sclerosis: comparing next-generation sequencing to sanger sequencing.*, 2014, volume 5, number 00166. The original manuscript is available at: doi: 10.3389/fneur.2014.00166

Introduction

Diagnosing diseases that affect the central nervous system (CNS) is inherently challenging. Multiple sclerosis (MS) is an autoimmune-mediated disease that exemplifies this challenge since clinicians must use multiple diagnostic tools to obtain the required evidence of dissemination of disease separated in time and space according to the current McDonald criteria (Polman et al., 2011). This includes radiological tests that detect lesions in the brain and spinal cord by magnetic resonance imaging (MRI) and is supported by biological tests that detect a unique pattern of oligoclonal banding (OCB) in the cerebrospinal fluid (CSF).

Due to the complexity associated with the current standard of care for MS diagnosis, patients who suffer an initial acute onset of “MS-like” symptoms (referred to as a clinically isolated syndrome - CIS) often have to wait before a diagnosis of MS is

confirmed and treatment is initiated (Milo and Miller, 2014). Steps to shorten this time frame are an urgent matter in the field, considering that patients have a better prognosis if treated early (Frohman et al., 2006a). Radiological testing (i.e. MRI) has been instrumental in the diagnosis of MS, but the most frequently used biological test that supports MS diagnosis is the OCB test, which has relatively low diagnostic specificity when comparing test performance for MS versus other neuro-inflammatory diseases (about 61%) (Reske et al., 2005; Tintore et al., 2008; Petzold, 2013).

The standardization of the OCB test to support a MS diagnosis led many neuroimmunologists in the field to focus on determining the role of B cells and their antibodies on the pathogenesis of MS (Lucchinetti et al., 2000; Cepok et al., 2005; Antel and Bar-Or, 2006; Owens et al., 2006; Harp et al., 2010). Early work by our group and others demonstrated that CSF-derived B cells from MS patients and CIS patients that convert to MS undergo extensive clonal expansion, skewing towards heavy chains of the 4^{th} family, and accumulate somatic hypermutations (SHM) at an advanced rate (Harp et al., 2007; Owens et al., 2007; Bennett et al., 2008). These features of antibody genetics are suggestive of a hyper-response to CNS antigens, but the targets of these CSF-derived B cells from MS patients remain elusive (Bennett and Owens, 2012). More recently, however, our laboratory has discovered that the 4^{th} family of heavy chain antibody genes of CSF-derived B cells from MS patients accumulates replacement mutations at 6 codon positions more frequently than patients with other neurological diseases (OND) (Cameron et al., 2009). B cells in MS lesions also display this pattern (Ligocki et al., 2010).

Using a custom algorithm to indicate the extent of mutation accumulation at these 6 codons in antibody gene repertoires, we developed a new biological test called the Antibody Gene Signature (AGS), which demonstrated promise in a small pilot cohort in identifying patients who had one demyelinating event and who would convert to MS (Cameron et al., 2009). However, these initial studies on the utility of AGS were based on Sanger sequencing, which is too laborious and expensive for routine use if this technology is developed as a clinical diagnostic test for MS in the future.

Next-generation DNA sequencing (NGS) might potentially provide a useful alternative in acquiring antibody gene repertoires to use for AGS calculations and is becoming routine in the field as evidenced by its commercial availability as a fee for service (Life Technologies, Illumina and SeqWright among many others). The most common application of NGS to antibody genetics has focused on VDJ recombination gene selection for the purpose of analyzing lymphocyte clonality (Boyd et al., 2009; Boyd et al., 2010; Arnaout et al., 2011; Logan et al., 2011), and is now being utilized in the MS field (von Budingen et al., 2012). Since gene and SHM distributions are at the core of antibody genetics analysis (as well as AGS scoring), careful scrutiny of this platform and its ability to properly represent the antibody gene repertoire is warranted.

Our primary goal was to provide confirmation that the antibody gene repertoires generated by NGS would sufficiently represent the CSF-derived B cell pool from MS patients. The data presented here demonstrate for the first time that antibody gene repertoires from individual CSF-derived B cells from the CSF of MS patients and those at high-risk to convert, generated by the gold standard Sanger method, are reliably reflected in NGS-generated antibody gene repertoires from paired CSF-derived B cell pools of the

same patients. Furthermore, we confirmed that AGS scoring, generated using a high-throughput NGS approach of pooled CSF cells, also identified MS patients and those that would convert to MS with the same accuracy as AGS scoring using Sanger DNA sequencing of individual CSF B cells. This NGS approach provides a new method for measuring the biological changes observed in MS patients and demonstrates its potential as a diagnostic tool.

Results

Sanger sequencing has been the gold standard to define the antibody repertoires of patients with autoimmune diseases such as MS (Owens et al., 1998; Qin et al., 1998; Baranzini et al., 1999; Colombo et al., 2000; Owens et al., 2001; Owens et al., 2003; Qin et al., 2003; Monson et al., 2005b). Such findings have provided necessary information to further our understanding on the role of B cells and their antibody products on the pathology of MS, the application of new targeting therapeutics, and the development of new diagnostic tools. NGS represents an advanced sequencing method to query even massive B cell pools, and has already been applied to defining B cell clonality in MS patients (von Budingen et al., 2012). However, it is critical to evaluate whether this new sequencing technology properly represents the unique features that were previously established by Sanger sequencing for antibody genetics in B cells from the CSF of MS patients.

Thus, we compared the antibody gene repertoires generated from single CSF-derived B cells using Sanger sequencing and those generated from CSF B cell pools using NGS in a cohort of MS/CIS patients. There were significant differences in the

frequency of individual VH4 gene usage between the platforms, although the relative abundance of individual *VH4* gene segments by rank was globally consistent (Figure 3-1A). In the comparison of the Sanger and NGS databases, *VH4-30*, *VH4-34* and *VH4-39* sequences show significant differences in abundance. *VH4-39* was the most abundant gene segment in the Sanger database, but is the third most abundant gene segment in the NGS database. All the other *VH4* gene segments remain in the same ranked order of abundance in both databases. The rank order of the *VH4-b*, *VH4-4* and *VH4-61* gene segments do not significantly vary between platforms. *VH4-59* has a significant increase in NGS (15% to 24%; p=0.004), which does not alter its rank. One noticeable difference is the lower abundance of long *VH4* gene segments (*VH4-30*, *VH4-31*, *VH4-39* and *VH4-61*) in the NGS database (23%) compared with the Sanger database (54%).

JH usage is important because skewing from the normal distribution of dominant JH4 usage (Briney et al., 2012) can be evidence of self-reactivity (Meffre et al., 2001). *JH4* remained the most abundant gene segment in both the Sanger and NGS databases (compare 38% to 40%; p=0.53) and *JH3* remained the fourth most abundant gene segment in both databases (compare 11% to 9%; p=0.18) (Figure 3-1B). *JH5* and *JH6* were significantly decreased in the NGS database, whereas *JH1* and *JH2* were significantly increased and resulted in significant differences in frequencies of these 4 JH genes between the platforms.

Skewing of mutation frequency and/or placement of mutations in antibody genes from the CSF of MS patients is well established (Monson et al., 2005b; Harp et al., 2007; Owens et al., 2007; Bennett et al., 2008). It is important, therefore, that the identification of the mutation accumulation and distribution is similar regardless of the platform by

which it was generated. With regard to the accumulation of mutations, the overall nucleotide mutation frequencies (MF) for individual patients by Sanger and NGS were similar (5.4% to 7.1%; p=0.16) (Figure 3-2A and Supplementary Table 3-1). The replacement mutation frequency (RMF) was also consistent between platforms (Figure 3-2B and Supplementary Table 3-1), again with a non-significant increase in NGS (9.7% to 12.5%; p=0.11). With regard to the distribution of mutations, the MF and RMF were also appropriately highest in the complementarity determining regions (CDRs), which are the antigen-contacting sites. The framework regions (FRs), which are the structural support regions of the antibody genes, had relatively few MF and RMF accumulations as expected (Figure 3-2A, 3-2B). The replacement to silent mutation ratios (R:S ratios) in the CDR regions increase from patient to patient (average 4.4 to 7.3; p=0.58) in the NGS platform, but without a significant trend emerging (Figure 3-2B). The R:S ratios in the FR regions were not significantly altered across platforms (1.4 to 1.5; p=0.94).

AGS scoring by Sanger sequencing showed initial success on a pilot cohort in identifying MS patients or CIS patients who will convert to MS (Cameron et al., 2009), which has been confirmed in larger sample cohorts (Figure 3-3A). To understand how antibody repertoire generation by NGS might affect AGS scoring calculations, we analyzed and compared the RMF at each codon position that defines the AGS (Figure 3-3B). Only codons 56 and 57 of the AGS maintained similar RMF to the Sanger repertoires. RMF at codons 40 and 89 were significantly increased and RMF at codons 31B and 81 were significantly decreased in comparison to the Sanger repertoires.

Despite these fluctuations in mutation distributions among the 6 AGS codons, we observed a non-significant change (14.9 to 12.2; p=0.22) in the paired samples of the

average AGS score with the NGS platform (Figure 3-3C)(Cameron et al., 2009). Two patients who have not yet received a confirmed RRMS diagnosis (patients C1 and C2) did not have consistent AGS scores between the Sanger and NGS databases (Figure 3-3D). However, all of those patients who did have RRMS or converted to RRMS after sampling showed consistent classification of disease by both Sanger sequencing and NGS. In addition, the specificity (50%), sensitivity (100%) and accuracy (85.7%) of properly identifying patients that have MS or would convert to MS in the future in this small cohort was the same for NGS-based, Sanger based, and oligoclonal banding biological testing. However, the small size of the cohort precludes any conclusion regarding the utility of NGS-based AGS scoring as a viable diagnostic test.

Finally, to understand these fluctuations in AGS scores between the two platforms, we show the distribution of AGS codon RM frequency and how it affects AGS scores for 3 representative samples. For example, in the Sanger repertoire of patient C2, approximately 21% of all RMs are within the AGS codons (Figure 3-3E) resulting in an AGS score of 13.07. In the NGS repertoire of this same patient, only 14% of all RMs are within the AGS codons resulting in a decreased AGS score of 4.43. Conversely, the NGS repertoire of patient C1 had an increased AGS score compared to the Sanger repertoire because of an increased percentage of RMs in AGS codons relative to all codons (compare 15% in Sanger vs 22% in NGS). Patient C4 had similar percentages of RM in AGS codons on both platforms (26% vs 25%), and thus had similar AGS scores on both platforms (17.90 vs 17.55).

Discussion

Radiological testing to support MS diagnosis has excelled and is indispensable in the diagnosis of MS, whereas development of biological tests to support MS diagnosis has been more challenging. One type of biological testing that is on the horizon is next generation sequencing (NGS), which can be used to query the antibody genetics of even massive B cell pools (Boyd et al., 2009; Boyd et al., 2010; Arnaout et al., 2011; Logan et al., 2011; von Budingen et al., 2012). Historically, this technology has been very successful in tracking minimal residual disease in cancer patients (Boyd et al., 2009). More recently the power of this technology has been used to demonstrate that focused B cell clones in the CSF of MS patients are identifiable in the vast peripheral B cell pools of the same patients (von Budingen et al., 2012). Thus, the use of NGS to pursue biological questions in MS has become a reality.

Our goal for this study was to advance beyond clonality queries and address whether the features of antibody genetics we had observed in CSF-derived B cells from MS patients with regard to antibody gene distribution (i.e. skewing towards *VH4* family usage) and somatic hypermutation accumulation (i.e. antibody gene signature) could be confirmed using this deeper sequencing method. This is important because NGS is now readily available commercially, and its possible limitations must be understood to best translate the information we obtain from it. To do this, we compared paired antibody repertoires generated from single CSF-derived B cells using Sanger sequencing and antibody repertoires generated from CSF B cell pools using NGS. This is the first time that there has been a direct comparison of this new technology to Sanger sequencing, which is the gold standard in the field.

Overall, we found that NGS and Sanger sequence data were similar with regard to general mutational profiles but differed somewhat in the distribution of *VH4* sub-family members recovered. Due to the similarity between the sequences of the *VH4* sub-family gene segments, the divergence in *VH4* distribution may be partially due to an increase in sequencing errors in the NGS database, the most common of which is insertion and deletion (indel) errors, particularly in regions that contain homopolymers or stretches containing 2 or more identical nucleotides (Bolotin et al., 2012). The reported frequency of indels generated by the Roche/454 platform is in the range of 3.8-5 x 10^{-3} (Loman et al., 2012; Georgiou et al., 2014). Indels are easily detected by alignment of NGS-generated sequences to published *VH4* sequences using the IMGT/High V-Quest tool (Alamyar et al., 2012). Since we remove all non-productive (with stop codons or frameshift mutations) or misaligned (<85% homology) antibody sequences, our NGS databases should contain very few sequences with indels. In order for a sequence with indels to pass our filters they would have to contain multiple complementary indel events in close proximity- an extremely unlikely scenario. Nucleotide substitution errors can also occur (Kircher and Kelso, 2010), but we used a very high-fidelity DNA polymerase to generate our NGS-based antibody repertoires so that the mutation frequencies between the Sanger and NGS databases would be similar.

All 5 patients who had or converted to RRMS were properly identified using our AGS biological test method by Sanger sequencing or NGS. There was some fluctuation in the AGS scores obtained for these paired samples between the two platforms, which could be due to a decreased representation in NGS of the long *VH4* genes that contain codon 31b. The AGS scoring system is based on mutation frequencies at 6 codons, which

includes 31b. Thus, if genes containing 31b are not properly represented in the NGS repertoire database in comparison to Sanger database, a decrease in AGS scores would be a natural consequence. Despite these differences in Sanger and NGS repertoire generation, identification of MS patients or CIS patients that would convert to MS remained the same between the two platforms.

The two CIS patients who did not convert to RRMS at follow-up are representative of biological testing complications due to patient care received. CIS patient C1 was at high risk to develop MS. The Sanger-based AGS score was below the 6.8 cut-off point, but the NGS-based AGS score was above the cut-off point suggesting that this patient would convert to RRMS in the future. CIS patient C2 was OCB positive at the time of sampling, with a single brain lesion noted by MRI, and was thus considered at low risk to develop MS. The Sanger-based AGS score was above the 6.8 cut-off point, but the NGS-based AGS score was below the cut-off point. In both of these cases, the patients were placed on disease modifying therapy shortly after sampling, making it difficult to determine what the natural progression of their demyelinating event may have been. Of note, patient C5 who was on steroids at the time of sampling and converted to RRMS had an AGS score above the 6.8 cut-off point by both platforms.

This study suggests that the transition from single B cell Sanger sequencing to high throughput NGS of pooled B cells is feasible with the application of appropriate sequence filtering methods to efficiently remove sequences containing errors generated during sample processing and sequencing. The implementation of appropriate quality metrics to identify and remove as many process-generated errors as possible will be critical for the successful use of NGS to better understand the antibody genetics of MS

and for the future development of a clinically useful NGS diagnostic test based on the AGS scoring algorithm. These results will need to be confirmed in a larger cohort of patients using NGS-based antibody repertoire generation before consideration as a diagnostic tool can be made.

Acknowledgements

The authors wish to thank the patients who provided samples for this study. We also thank Dr. Lee Szkotnicki and SeqWright Genomic Services (a GE company, Houston, TX) for performing PCR amplification and sequencing to generate the NGS database. We also want to thank Dr. Mark Erlander (Trovagene, San Diego, CA) for his critical reading of the manuscript. This study was supported by grants from the National Multiple Sclerosis Society (NMSS) to NLM (RG3267 and RG4653) and DioGenix, Inc. to NLM. AJL and WHR were supported by Grant no. NIH NRSA5 T32 A1 005284-28 from NIAID.

FIGURE LEGENDS FOR CHAPTER THREE: RESULTS

Figure 3-1. ***VH4*** **gene distributions show cross-platform variation for samples from both patients with RRMS and CIS.** *VH4* (A) and *JH* (B) gene calls were obtained by IMGT alignment. Total sequences used in Sanger sequencing and next-generation sequencing (NGS) databases are indicated inside the pie charts. Statistically significant differences between the frequencies of individual genes were identified by Chi-squared test (p-value: N.S. ≥ 0.05).

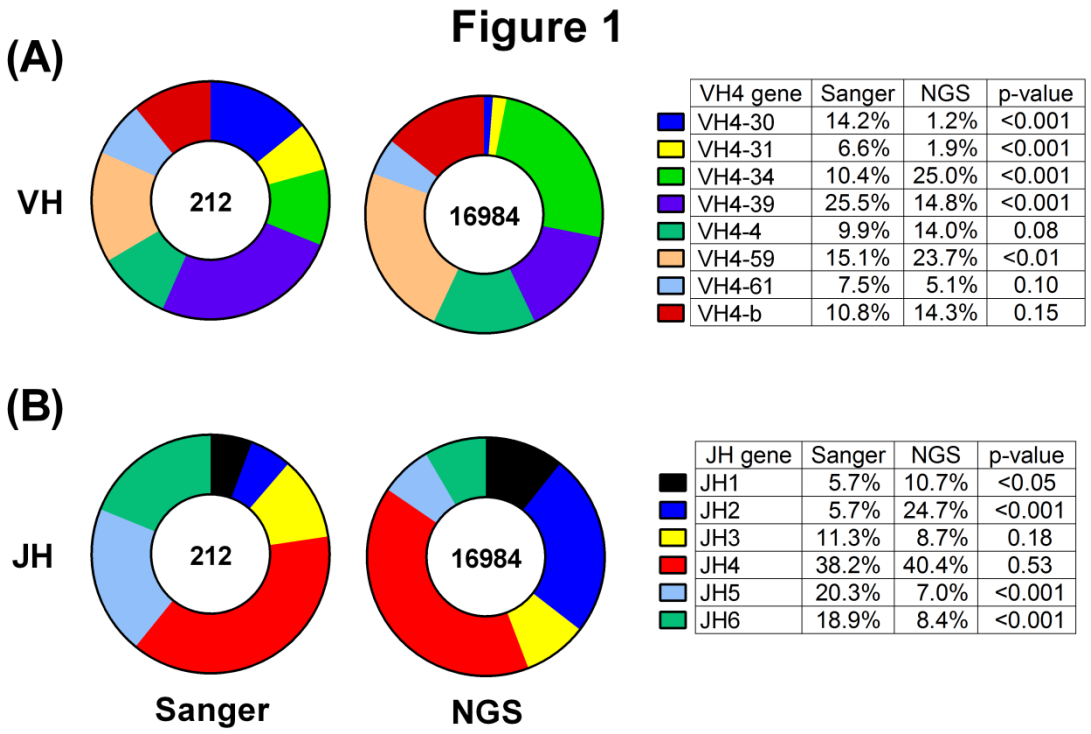
Figure 3-2. Mutation characteristics of ***VH4*** **sequences in RRMS and CIS patients.** Sanger sequence data includes 212 sequences with 2265 total point mutations and 1386 total replacement mutations (RM). Next-generation sequencing (NGS) data includes 16,984 unique sequences with 263,764 total point mutations and 154,457 total replacement mutations (RM). (A) Mutation frequency (MF) analysis was done by nucleotide; (B) Replacement mutation frequency (RMF) analysis was done by codon. MF and RMF were calculated by patient and bar graphs show mean (indicated on the bar graphs) and S.D. (statistical significance of the distributions was tested for by Wilcoxon matched-pairs signed rank test; N.S. ≥ 0.05). MF, RMF and R:S ratios for CDR and FR regions were calculated independently by region for each patient and are shown as patient means.

Figure 3-3. Antibody Gene Signature (AGS) in RRMS and CIS patients. (A) Unpaired Sanger sequence datasets for multiple sclerosis (MS, includes relapsing-remitting, primary and secondary progressive MS samples) and other neurological

disease (OND) cohorts. Each data point represents a single patient sequence pool that was not analyzed by NGS. The dotted line represents the AGS cut-off point of 6.8 above which patients are expected to convert to relapsing-remitting multiple sclerosis (RRMS). Mean and standard deviation are shown. (B) Replacement mutation frequency (RMF) of each of the 6 AGS codons was calculated relative to the total AGS RM in each dataset. P-values were calculated by Chi-squared test. (C) Each data point represents a single patient sequence pool. The dotted line represents the AGS cut-off point of 6.8 above which patients are expected to convert to RRMS. Mean and standard deviation are shown. Statistical significance of the distributions was tested for by Wilcoxon matched-pairs signed rank test (N.S. ≥ 0.05). (D) The AGS scores of the 7 paired patients are shown here. (E) The percent of total RMs that belong to the AGS pattern in each sequence was mapped for 3 patients with different types of AGS score shifts from one platform to another. The boxes indicate mean and the error bars S.D.

FIGURES FOR CHAPTER THREE: RESULTS

Figure 3-1.



VH4 gene	Sanger	NGS	p-value
VH4-30	14.2%	1.2%	<0.001
VH4-31	6.6%	1.9%	<0.001
VH4-34	10.4%	25.0%	<0.001
VH4-39	25.5%	14.8%	<0.001
VH4-4	9.9%	14.0%	0.08
VH4-59	15.1%	23.7%	<0.01
VH4-61	7.5%	5.1%	0.10
VH4-b	10.8%	14.3%	0.15

JH gene	Sanger	NGS	p-value
JH1	5.7%	10.7%	<0.05
JH2	5.7%	24.7%	<0.001
JH3	11.3%	8.7%	0.18
JH4	38.2%	40.4%	0.53
JH5	20.3%	7.0%	<0.001
JH6	18.9%	8.4%	<0.001

Figure 3-2.

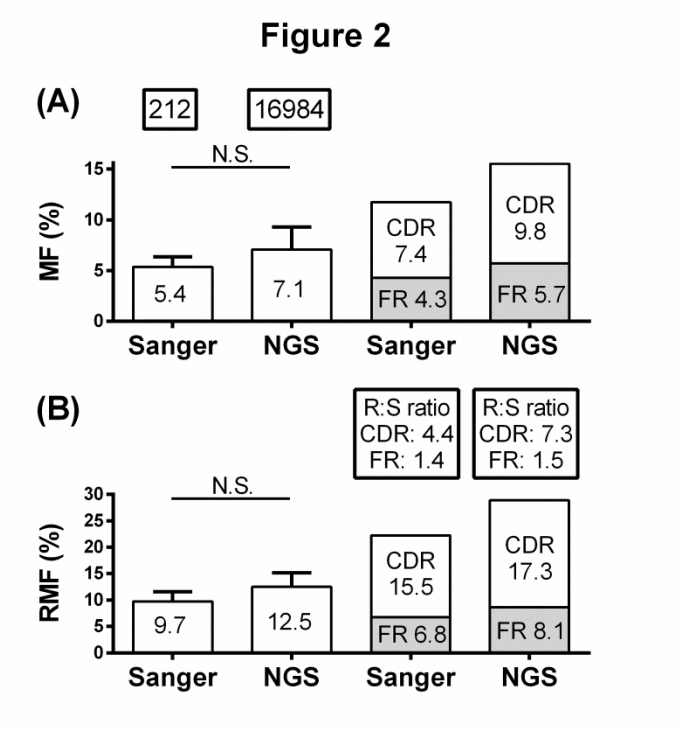
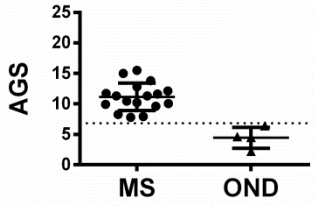


Figure 2
(A)
212
16984
N.S.
MF (%)
15
10
5
0
5.4
7.1
CDR 7.4
FR 4.3
CDR 9.8
FR 5.7
Sanger
NGS
Sanger
NGS
(B)
R:S ratio CDR: 4.4 FR: 1.4
R:S ratio CDR: 7.3 FR: 1.5
N.S.
RMF (%)
30
25
20
15
10
5
0
9.7
12.5
CDR 15.5
FR 6.8
CDR 17.3
FR 8.1
Sanger
NGS
Sanger
NGS

Figure 3-3.

Figure 3

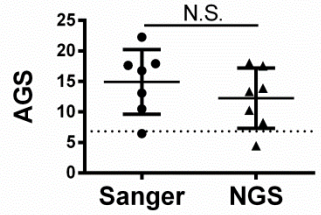
(A)



(B)

AGS codon	Location	Sanger RMF	NGS RMF	Fold increase	p-value
31B	CDR1	17.04%	2.87%	0.17	0.001
40	FR2	13.18%	28.92%	2.19	0.001
56	CDR2	25.08%	26.05%	1.04	0.700
57	CDR2	9.00%	6.50%	0.72	0.075
81	FR3	27.97%	18.80%	0.67	0.001
89	FR3	7.72%	16.86%	2.19	0.001

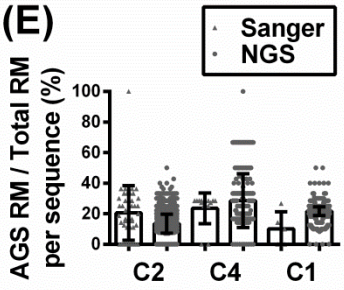
(C)



(D)

patient	Sanger AGS	NGS AGS	Current Diagnosis
C1	6.43	13.32	CIS
C2	13.07	4.43	CIS
C3	10.47	13.88	RRMS
C4	17.90	17.55	RRMS
C5	16.73	8.21	RRMS
C6	17.62	10.26	RRMS
C7	22.26	18.01	RRMS

(E)



TABLES FOR CHAPTER THREE: RESULTS

Table 3-1. Patient sample summary. Initial diagnosis at the time of sample collection is indicated for each patient in the study. Abbreviations in table: OCB, oligoclonal bands; AGS, antibody gene signature; CIS, clinically isolated syndrome; RRMS, relapsing-remitting multiple sclerosis.

Patient ID	Initial diagnosis[1]	OCB status	Comments	Follow-up diagnosis[2]	Follow-up time[3]	Age[4]	Gender	Sanger AGS	NGS AGS
C1	CIS	NEG	High risk of RRMS	CIS	44	45	F	6.43	13.32
C2	CIS	POS	Single lesion[5]	CIS	26	34	F	13.07	4.43
C3	CIS	POS		RRMS	1	39	F	10.47	13.88
C4	CIS	POS	High risk of RRMS	RRMS	8	27	F	17.90	17.55
C5	RRMS	POS	On steroids	RRMS	36	19	F	16.73	8.21
C6	RRMS	POS		RRMS	25	19	F	17.62	10.26
C7	CIS	POS	High risk of RRMS	RRMS	31	33	M	22.26	18.01
C8	CIS	POS	Low risk of RRMS	RRMS	8	34	F	10.17	NA

[1] At time of sampling using 2005 McDonald Criteria
[2] Using 2010 McDonald Criteria
[3] Since sampling (months)
[4] At time of sampling (yrs)
[5] by MRI of the brain

Table 3-2. Sequence database size summary. For each patient, the initial *VH4* sequences obtained by Sanger sequencing of single B cells, the number of B cells in the cell pellet used for NGS PCR and sequencing and the number of unique *VH4* NGS sequences after filtering are indicated. Of note, a typical Sanger-based antibody repertoire can take several months to generate whereas NGS-based repertoires can take as little as one week.

Patient ID	# of Sanger *VH4* sequences	# of B cells in cell pellet for NGS	# of unique NGS *VH4* sequences
C1	7	29	2,475
C2	41	100	2,213
C3	61	100	14
C4	14	30	596
C5	25	100	5,020
C6	46	100	4,290
C7	18	100	2,376
Ave.	30		2,426

Table 3-3. (Supplementary) Mutation characteristics of *VH4* sequences in RRMS and CIS patients. Mutation frequency (MF) analysis was done by nucleotide. Replacement mutation frequency (RMF) analysis was done by codon. MF and RMF means were calculated by patient and statistical significance of the frequency distributions between Sanger and NGS databases was tested for by Wilcoxon matched-pairs signed rank test be.

Patient ID	MF		RMF	
	Sanger	NGS	Sanger	NGS
C1	4.75%	7.73%	8.42%	13.84%
C2	6.61%	9.53%	12.38%	14.70%
C3	5.83%	4.93%	10.69%	9.28%
C4	6.42%	3.49%	11.43%	8.12%
C5	5.00%	8.47%	8.79%	14.64%
C6	3.74%	6.52%	7.18%	13.37%
C7	5.15%	8.80%	9.07%	13.63%
Wilcoxon test	0.156		0.109	

CHAPTER FOUR

RESULTS

AIM I: AGS scoring by next-generation sequencing is a reliable replacement for MS conversion diagnosis by single-cell Sanger sequencing analysis

Overview and rationale

After showing that next generation sequencing is an efficient method for obtaining the sequencing information required AGS scoring, we wanted to test its performance on CSF cell pellets, since these are easier to collect than sorted CD19+ B cells. As a result, the focus of this "Verification" study was primarily to test AGS performance using a new bulk cell gDNA extraction protocol, as opposed to the methods used in the previous study (Chapter 3; Table 2-1). Additionally, the number of OND samples included was increased to match RRMS patient numbers.

MSPRECISE: A MOLECULAR DIAGNOSTIC TEST FOR MULTIPLE SCLEROSIS USING NEXT GENERATION SEQUENCING.

The following study has been published in *Gene*. Rounds WH, Salinas EA, Wilks TB, 2nd, Levin MK, Ligocki AJ, Ionete C, Pardo CA, Vernino S, Greenberg BM, Bigwood DW, Eastman EM, Cowell LG, Monson NL, *MSPrecise: A molecular diagnostic test for multiple sclerosis using next generation sequencing.*, 2015, volume 572, issue 2, pages 191-197. It is reproduced here with permission from Elsevier for use in a printed thesis, with the original manuscript available at: doi:10.1016/j.gene.2015.07.011

Introduction

Multiple sclerosis (MS) is a demyelinating autoimmune disease of the central nervous system (CNS). Several studies have underscored the impact of T and B cells in this disease and have broadened the community's search for more effective immunomodulatory therapies for the treatment of relapsing-remitting MS (RRMS). For example, early evidence for a role of B cells in the pathoetiology of MS, including oligoclonal bands,(Andersson et al., 1994; Krumbholz et al., 2012) altered antibody genetics (Harp et al., 2007; Owens et al., 2007; Bennett et al., 2008) and B cell responses to neuroantigens in vitro (Antel and Bar-Or, 2006; Harp et al., 2007) provided the basis for use of Rituximab, a B cell depleting antibody for the efficacious treatment of RRMS.(Hauser et al., 2008; Kappos et al., 2011)

A number of reports consistently demonstrate that B cells in the CNS of RRMS patients undergo extensive clonal expansion,(Qin et al., 1998; von Budingen et al., 2008; Krumbholz et al., 2012; von Budingen et al., 2012) and in some cases, recognize neuroantigens. Our laboratory hypothesized that since antigen-driven B cell selection is dependent on somatic hypermutation (SHM) accumulation in antibody genes, the cerebrospinal fluid (CSF)-derived B cell pool of RRMS patients would be enriched for a unique pattern of SHM reflecting their potential to recognize neuroantigens. Since variable heavy chain family 4 (VH4) genes are enriched in RRMS patient CNS,(Owens et al., 1998; Baranzini et al., 1999; Colombo et al., 2000; Owens et al., 2003; Monson et al., 2005b; Harp et al., 2007; Owens et al., 2007) this gene family was examined for patterns of SHM. Indeed, we have demonstrated and confirmed that CSF-derived B cells from RRMS patients expressing rearranged variable heavy chain family 4 (VH4) genes have an exaggerated accumulation of replacement mutations at 6 codon positions.(Cameron et al., 2009; Rounds et al., 2014)

Earlier studies of this SHM pattern used a pool of memory B cells isolated from healthy donor peripheral blood (N=2) to establish baseline SHM accumulation at each codon position. Our next goal was to compare the SHM pattern identified in MS patients with CSF B cell antibody repertoires from patients with other neurological diseases (OND). However, these early studies included comparison to only 3 OND patients. Thus, further confirmation is required regarding the specificity of SHM accumulation at these codon positions in B cells from RRMS patients and a larger OND cohort. In addition, the majority of patients analyzed in the previous two studies were patients who were very early in their disease (N=17/19). Thus, it is unclear whether established RRMS patients

who meet the McDonald criteria for RRMS (Polman et al., 2011) have the same exaggerated accumulation of SHM at these codon positions.

To address these issues, we analyzed the VH4 antibody gene repertoires in CSF cell pellets from 26 patients with OND and 13 patients with confirmed RRMS using next generation sequencing (NGS). Our results indicate that RRMS patients exhibited the expected pattern of SHM at these codon positions. In addition, 23/26 OND patients did not appreciably accumulate SHM at these codon positions or displayed insufficient sequence data indicative of low B cell abundance in the CSF.

Results

For this study, we generated *VH4* antibody repertoires using NGS of CSF cell pellets isolated from 39 patients (Table 4-1). Of the 39 patient-derived CSF cell pellets, 13 were from patients with confirmed or possible RRMS, and 26 were from patients with OND. 14 patient samples (1 RRMS and 13 OND) were excluded due to recovery of insufficient sequence reads after sequence filtering (Tables 4-2 & 4-3). A pool of purified CD19+CD27- naïve B cells from peripheral blood of one healthy donor (run in 10 replicates) was included as a sequencing control for 454 error rates and as a control for random *VH4* gene usage in the naïve B cell pool.

We first determined how a series of process and analytical modifications made since previous analyses affected sequence coverage (Supplementary Methods 1.1 and 1.2).(Rounds et al., 2014) One modification was to include only unique sequences that had two or more copies after sequence filtering (redundancy 1) in an attempt to increase our confidence that the sequences being analyzed were representative of the B cell pool

and not a result of sequence errors generated during either PCR amplification or NGS. We compared the sequence coverage obtained with redundancy filter (R1) and without (R0) (Table 4-4). The previously published dataset had an average of 2,426 unique sequences per RRMS sample at R0 and an average of 583 sequences per RRMS sample at R1. The current dataset had an average of 751 sequences for the RRMS samples and an average of 632 sequences for the OND samples at R1 (Table 4-4). This resulted in a 1.3-fold increase per RRMS patient in the number unique sequences in CSF-derived antibody repertoires using our current method. The healthy control naïve (HCN) cohort had an average of 1,363 sequences per sample, which resulted in 2.5-fold more coverage in the peripheral HCN B cell pools in comparison to all CSF B cell pools, which likely relates to a larger initial pool of purified B cells.

Next, we sought to determine if the distributions of variable heavy chain family 4 (*VH4*) gene segments in each cohort were comparable (Figure 4-1a). The *VH4* gene distributions differed significantly between all pairs of cohorts with some pairs being more divergent than others. The RRMS *VH4* gene distribution was most distinct relative to the other two cohorts (Chi-squared value = 5652 for RRMS versus HCN; 3741 for RRMS versus OND), while the OND and HCN distributions were more similar (Chi-squared value = 2114). As expected,(Brezinschek et al., 1995) the usage frequency of *VH4* genes in the HCN B cell pool was comparable to a uniform distribution of 12.5% for each individual gene (Chi-squared value = 4665), with an underrepresentation of *VH4-4* (percentage deviation = -81%) and an overrepresentation of *VH4-b* (percentage deviation = 119%) contributing most to the overall Chi-squared value. Similarly, for the OND cohort, deviation from a uniform distribution of gene usage is primarily due to one or two

genes, with underrepresentation of *VH4-31* showing the largest deviation (percent deviation = -96%). In contrast, the RRMS cohort was very different from a uniform distribution (Chi-squared value = 7804) and utilized *VH4-39* (percentage deviation = 190%) and *VH4-59* (percentage deviation = 105%) more frequently than expected, which others have previously observed for *VH4-39*.(Owens et al., 1998; Baranzini et al., 1999)

The distribution of joining heavy chain (JH) gene segments in naïve B cells is heavily skewed towards *JH4* usage.(Brezinschek et al., 1995) Indeed, the healthy donor peripheral naïve B cell pools in the current dataset demonstrated skewing towards *JH4* usage (Figure 4-1b). However, the RRMS cohort for this dataset had a JH usage rank of 5>6>4>2>1=3. The high usage of *JH5* and *JH6* gene segments was unexpected and contrasted with the previous dataset where *JH4* was maintained as the most frequently used JH gene segment in the RRMS cohort.(Rounds et al., 2014) Further investigation confirmed that 8 of the 12 RRMS patients had unusually high skewing towards *JH5* or *JH6* usage, which resulted in an unexpected JH usage rank in the cohort. Thus, the overall distribution of JH gene segments in the RRMS cohort was significantly different from that of the HCN cohort (Chi-squared value = 2416). The OND cohort had a JH gene segment usage rank of 4>5>6>1=3>2, which more closely followed the JH rank of the HCN B cell repertoire (Chi-squared value = 1791).

We next determined whether the RRMS and OND cohorts from this dataset had accumulated SHMs into the variable regions of their antibody genes as established in the literature by calculating both the overall mutation frequency (MF), which considers all nucleotide substitutions, and the replacement mutation frequency (RMF), which considers only amino acid substitutions (Figure 4-2). Whereas the HCN B cell pools had

very low MFs (median 1.9%) as expected from a naïve B cell population with low background sequencing error, the RRMS and OND cohorts had very high MFs (medians 6.7% for RRMS and 3.4% for OND), demonstrating that CSF B cells accumulate SHMs at a high frequency as previously published.(Monson et al., 2005a) Interestingly, the MF of the RRMS and OND cohorts were not significantly different (p=0.50). The RMF calculations demonstrate a similar result (i.e. high and comparable RMF in the RRMS and OND CSF cohorts compared to the peripheral HCN). No correlation was found between patient age and RMF for either cohort (RRMS p=0.8; OND p=0.2). Proper targeting of these mutations to the hypervariable regions within the complementarity determining regions (CDRs) was also confirmed (Figure 4-2b).

Next, we compared the RMF at each codon position in the 6 codons that we originally used to calculate antibody gene signature (AGS) scores (31B, 40, 56, 57, 81, 89)(Cameron et al., 2009). The RMF at codons 31B, 40, 56, and 57 were all statistically greater in the RRMS cohort compared to the OND cohort (Table 4-5). However, the RMF at codons 81 and 89 were statistically greater in the OND cohort compared to the RRMS cohort. In fact, codon 89 had the lowest RMF of all 6 AGS codons in the RRMS cohort (9.3%), and thus contributed the least to scores for the RRMS cohort combined as well as for individual patients.

Finally, we calculated **MS*Precise*** scores for all 25 patient CSF samples (Figure 4-3), excluding codon 89 in the calculations due to its low impact on scores for the RRMS cohort. As expected, the RRMS samples had a median **MS*Precise*** score of 10.6 and IQR of 5.7 to 17.7. The OND samples had a median **MS*Precise*** score of 4.5 and IQR of -3.3 to 11.7. Thus, the **MS*Precise*** scores of the RRMS cohort were statistically higher than

the **MS***Precise* scores of the OND cohort (p=0.05). The HCN cohort had very consistent and low **MS***Precise* scores as expected for a sequencing control that demonstrates non-targeted background sequence error, with a median score of -0.6 and an interquartile range (IQR) of -1.1 to 0.6.

As expected, 10 of 13 OND patients had **MS***Precise* scores below the previously established threshold of 6.8. However, the 6.8 threshold was based on Sanger sequencing data and NGS sequences have a low level of background RMs which tends to lower **MS***Precise* scores. Therefore, we identified an alternative threshold of 5.8 where we would expect to find some NGS samples with **MS***Precise* scores above but close to the threshold by Sanger sequencing. This new threshold did not affect the number of OND patients that had **MS***Precise* scores low enough to be properly identified. Four of the OND patients had **MS***Precise* scores just below the 5.8 **MS***Precise* threshold, and six of them had negative **MS***Precise* scores. There was no correlation between diagnoses of the OND patients and their **MS***Precise* scores (Tables 4-2 & 4-3).

MS*Precise* scores for 9 of the 12 RRMS patients were above the **MS***Precise* threshold of 5.8 and included 2 patients who were on interferon beta-1a (one for 9 months, MS05, and one for 2 years, MS07), one patient who was on glatiramer acetate for 5 years (MS04) and one patient who was on mycophenolic acid for 7 years (MS06). All four of the patients diagnosed with RRMS who were oligoclonal banding (OCB) negative had **MS***Precise* scores above the threshold (scores = 33.2; 10.0; 26.8; 7.5), two of which were on disease-modifying therapies (DMT) (MS06, MS07). Of the three RRMS patients who were OCB positive, two had **MS***Precise* scores below the threshold (scores = -3.5 and 6.6), but had been sampled while on DMTs (MS02, steroids; MS05,

interferon beta-1a). One OCB positive RRMS patient who was not on DMT at the time of sampling had an **MS*Precise*** score above the threshold (score = 15.2).

No correlations were found between **MS*Precise*** score and age or mutation frequency (Figure 4-5). There was a trend towards higher diversity in *VH4* gene usage (termed "diversity index") for RRMS patients with low **MS*Precise*** scores (Figure 4-4a), which did not correlate with sequence read count (Figure4-6). The two RRMS patients that had high diversity indices and low **MS*Precise*** scores were MS08 (diversity index = 1.10; score = 5.37) and MS10 (diversity index = 1.22; score = -1.51). The OND cohort did not display any correlation of **MS*Precise*** score with the diversity index (Figure 4-4b), even though the diversity index for the RRMS and OND cohorts were not statistically different (Figure 4-4c; $p=0.6$). The HCN cohort displayed a high diversity index that was statistically different from both the RRMS and OND cohorts ($p<0.0001$ for both) as expected from a large peripheral B cell pool compared to CSF B cells (Figure 4-4c).

Discussion

The application of antibody genetics to human disease has begun to emerge rapidly, particularly since NGS became readily available. Indeed, the power of this technology has been applied to monitoring minimal residual disease in cases of B cell lymphomas (Boyd et al., 2009), and establishing that CSF-derived B cell clones matriculate from the periphery.(von Budingen et al., 2012) Our application of NGS has been to develop a new approach to identify patients with clinically isolated syndrome (CIS) who are at high risk of converting to fulminant MS. Indeed, our early work using

Sanger DNA sequencing methods demonstrated that AGS scoring identified CIS patients who later converted to definite RRMS with 91% accuracy.(Cameron et al., 2009)

However, four questions remained. First, was the accumulation of SHM in these codons specific to MS patients? Second, would established RRMS patients that meet the revised McDonald criteria (Polman et al., 2011) have a similar pattern of SHM as early-stage patients? Third, does OCB status affect the score? Fourth, does treatment with immunomodulatory drugs affect the score? To address these issues, we generated antibody gene repertoires from CSF-derived B cells of ONDs, OCB^+ and OCB^- RRMS patients as well as treatment-naïve RRMS patients and RRMS patients who had been on DMTs for more than a year.

We obtained CSF cell pellets from 26 OND patients with a variety of diagnoses including headache (n=6), paraneoplastic disease (n=4) and others (Tables 4-2 & 4-3). Of the 26 OND patients, 13 were excluded from analysis due to a very low number of sequence reads. Since this primarily occurred in the OND cohort, we concluded that those 13 OND patients either did not display an expanded B cell mediated CNS immune response that we could detect, or that the response was negligible. In either case, the inability to recover antibody sequences from such samples is likely indicative of a lack of B cell recruitment and confirms why the literature is limited in the area of antibody genetics in patients with non-inflammatory neurological diseases. In fact, there was one RRMS patient with insufficient reads that we did not include in the present cohort because this patient had been on natalizumab for more than 4 years, a well-known drug that prevents B cells and other lymphocytes from entering the CNS.(Stuve and Bennett, 2007)

Our ability to detect antibody genes of rare B cells by PCR might provide OND samples an advantage and result in an **MS***Precise* score that might not properly reflect their OND status. In addition, low antibody sequence reads might be indicative of their OND status. Indeed, of the 14 samples we removed based on recovery of an insufficient number of unique sequence reads, 13 of them were within the OND cohort. If we assigned such samples the lowest **MS***Precise* score possible (**MS***Precise* score = -8.9), and inserted them back into the OND cohort, the median **MS***Precise* score of the OND group decreases to -8.9 (Figure 4-7).

In those 13 OND cases where we were able to recover a sufficient number of unique antibody sequences from CSF-derived cells, we observed that the accumulation of replacement mutations was slightly lower than in the RRMS patients, but not significantly different (OND, median RMF 6.5; RRMS, median RMF 9.9; p=0.5). In addition, the distribution of *VH4* gene segments in the OND cohort did not differ significantly from the expected random frequency. JH gene segment usage was also no different from the expected frequency established in naïve B cell pools. This suggests that in the OND cases for which CSF B cells can be detected, antigen-driven selection is not as prominent as it is in RRMS patients.

There is very little available information regarding the impact of DMTs on numbers or types of B cells found in the CSF of RRMS patients. Even in the case of B cell-depleting monoclonal antibodies, such as Rituximab, our understanding of B cell dynamics in the CSF is limited.(Monson et al., 2005a; Evdoshenko et al., 2013) Nevertheless, the RRMS cohort used for this study included 4 patients on DMTs for an extended period of time, most of which had high **MS***Precise* scores regardless of OCB

status. The one RRMS patient who had been on steroids for 7 days at the time of sampling had a negative **MS***Precise* score. It is difficult to make conclusions based on these small samples, but these data suggest that the clinical benefit of many immunomodulatory drugs used to treat RRMS, including the beta-interferons and glatiramer acetate, is independent of the CSF B cell pool. Further study is warranted to determine if particular DMTs impact the CSF B cell pool and **MS***Precise* scores.

Finally, there is an increasing need for new methods to determine whether a patient has MS or not.(Kroksveen et al., 2014) **MS***Precise* scoring may be one supportive approach to aid clinicians in this task. Indeed, if we include the OND samples with insufficient reads, the specificity of identifying patients with OND based on **MS***Precise* scoring is 88%. The sensitivity of this test in identifying RRMS patients is 75%, although the impact of DMTs and steroids on the **MS***Precise* scoring system for our RRMS cohort remains unclear. This puts the overall accuracy of **MS***Precise* scoring in this study at 84% if samples with insufficient reads are included and 76% if they are omitted. Previously, we presented data generated using Sanger DNA sequencing suggesting that **MS***Precise* scoring is able to identify CIS patients who will convert to RRMS but who are not yet on immunomodulatory therapy with 91% accuracy.(Cameron et al., 2009) Determining whether **MS***Precise* scoring using NGS performs as well to identify CIS patients who will convert to RRMS will be the subject of future investigations. More work also needs to be done to determine whether the codons we used to calculate **MS***Precise* scores are still appropriate on the NGS platform which will require a larger patient cohort with preferably several sub-cohorts of RRMS patients on particular DMTs and OND patients of a particular diagnosis.

Acknowledgements

The authors wish to thank the patients who provided samples for this study. We thank Carolyn Griffin for coordinating sample collection at the University of Massachusetts. We also thank Dr. Lee Szkotnicki and SeqWright Genomic Services (a GE company, Houston, TX) for performing PCR amplification and sequencing to generate the NGS database. We also want to thank Dr. Mark Erlander (Trovagene, San Diego, CA) for his critical reading of the manuscript.

FIGURE LEGENDS FOR CHAPTER FOUR: RESULTS

Figure 4-1. ***VH4*** **and JH gene distributions of CSF B cells from RRMS patients are more divergent from healthy control naïve peripheral B cell repertoires than those from OND patients.** *VH4* (a) and JH (b) gene calls were obtained by IgBlast alignment (see methods). Total unique sequences used in cohort databases are indicated inside the pie charts. Chi-squared analysis values between cohort gene distributions are shown above the bars. Gene frequencies are shown in the table. Abbreviations: RRMS, relapsing-remitting MS; OND, other neurological disorder; HCN, healthy control naïve peripheral B cells. HCN samples are all replicates from a single patient.

Figure 4-2. Mutation characteristics of ***VH4*** **sequences in RRMS and OND patients.** (a) Mutation frequency (MF) analysis was done by nucleotide; boxes indicate total unique sequences in each cohort and sample numbers are marked under cohort names. (b) Replacement mutation frequency (RMF) analysis was done by codon. RRMS sequence data includes 119,483 total point mutations and 62,749 total replacement mutations (RM); OND sequence data includes 74,769 total point mutations and 39,324 total replacement mutations (RM); RRMS sequence data includes 51,238 total point mutations and 17,375 total replacement mutations (RM). MF and RMF were calculated by sample and bar graphs show median (indicated on the bar graphs) and interquartile range (statistical significance of the difference between RRMS and OND was tested by Mann Whitney test). MF, RMF and R:S ratios for CDR and FR regions were calculated independently by region for each sample and are shown as cohort medians. HCN samples are all replicates from a single patient.

Figure 4-3. MS*Precise* scores in RRMS and OND patients. Each data point represents a single sample sequence pool (median and interquartile range are marked on the figure). The dashed line represents the **MS*Precise*** cut-off point of 6.8 above which patients are expected to have or convert to relapsing-remitting MS (RRMS). The dotted line delineate an indeterminate range (-1) below the 6.8 cut-off where the results of the **MS*Precise*** score test are less clear cut. Samples are grouped by most current diagnosis as RRMS, other neurological diseases (OND), and healthy control naïve (HCN). Only samples that pass our filtering criteria are displayed with their calculated **MS*Precise*** scores. Statistical significance of the difference between cohorts was calculated by Mann Whitney test. HCN samples are all replicates from a single patient.

Figure 4-4. Low diversity correlates with high MS*Precise* score in the RRMS cohort but not in the OND cohort. Each data point represents a single sample sequence pool from (a) the RRMS cohort or (b) the OND cohort. The diversity index was calculated as described in the methods section and high values indicate a more even distribution across the *VH4* genes. Pearson's correlation coefficient (R) indicates the linear correlation between **MS*Precise*** and the diversity index, and the two-tailed p-value of the correlation is also indicated. The dashed line represents the **MS*Precise*** cut-off point of 6.8 above which patients are expected to have or convert to relapsing-remitting MS (RRMS). The dotted lines delineate an indeterminate range (-1) below the 6.8 cut-off where the results of the **MS*Precise*** score test are less clear cut. (c) Distribution of the diversity index is shown here with the median marked on the graph. HCN samples are all replicates from a

single patient. Statistical significance of the difference between cohorts was tested by Mann Whitney test.

Figure 4-5. MSPrecise score does not correlate with age, MF% or RMF% in both RRMS and OND. Each data point represents a single sample sequence pool. Pearson's correlation coefficient (r) indicates the linear correlation between MSPrecise and either age in years, mutation frequency (MF%), or replacement mutation frequency (RMF%). The two-tailed p-value of the correlation is also indicated.

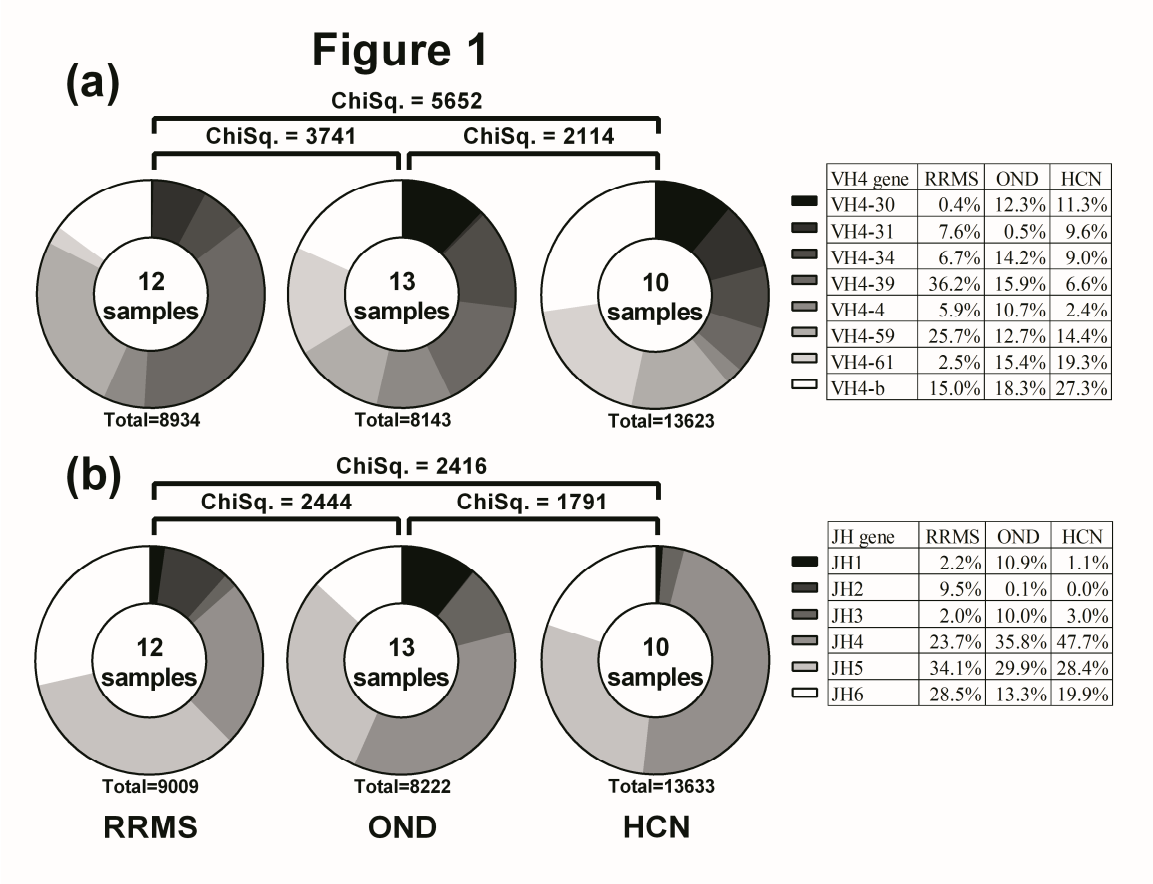
Figure 4-6. Diversity index does not correlate with sequence number in both RRMS and OND. Each data point represents a single sample sequence pool. 2 high sequence number outliers were removed because they had more than the median + 2 standard deviations of the sequences of all CSF samples (> 1,431 unique sequences). Pearson's correlation coefficient (r) indicates the linear correlation between the diversity index and the number of unique sequences in the sample. The two-tailed p-value of the correlation is also indicated.

Figure 4-7. MSPrecise scores in all RRMS and OND patients. Each data point represents a single sample sequence pool (median and interquartile range are marked on the figure). The dashed line represents the MSPrecise cut-off point of 6.8 above which patients are expected to have or convert to relapsing-remitting MS (RRMS). The dotted line delineates an indeterminate range (-1) below the 6.8 cut-off where the results of the MSPrecise score test are less clear cut. Samples are grouped by most current diagnosis as RRMS, other neurological diseases (OND), and healthy control naïve (HCN). OND

samples that were filtered out due to low sequence count are added to with an assigned MSPrecise score of -8.9 (minimum score). Statistical significance of the difference between cohorts was tested by Mann Whitney test.

FIGURES FOR CHAPTER FOUR: RESULTS

Figure 4-1.



VH4 gene	RRMS	OND	HCN
VH4-30	0.4%	12.3%	11.3%
VH4-31	7.6%	0.5%	9.6%
VH4-34	6.7%	14.2%	9.0%
VH4-39	36.2%	15.9%	6.6%
VH4-4	5.9%	10.7%	2.4%
VH4-59	25.7%	12.7%	14.4%
VH4-61	2.5%	15.4%	19.3%
VH4-b	15.0%	18.3%	27.3%

JH gene	RRMS	OND	HCN
JH1	2.2%	10.9%	1.1%
JH2	9.5%	0.1%	0.0%
JH3	2.0%	10.0%	3.0%
JH4	23.7%	35.8%	47.7%
JH5	34.1%	29.9%	28.4%
JH6	28.5%	13.3%	19.9%

Figure 4-2.

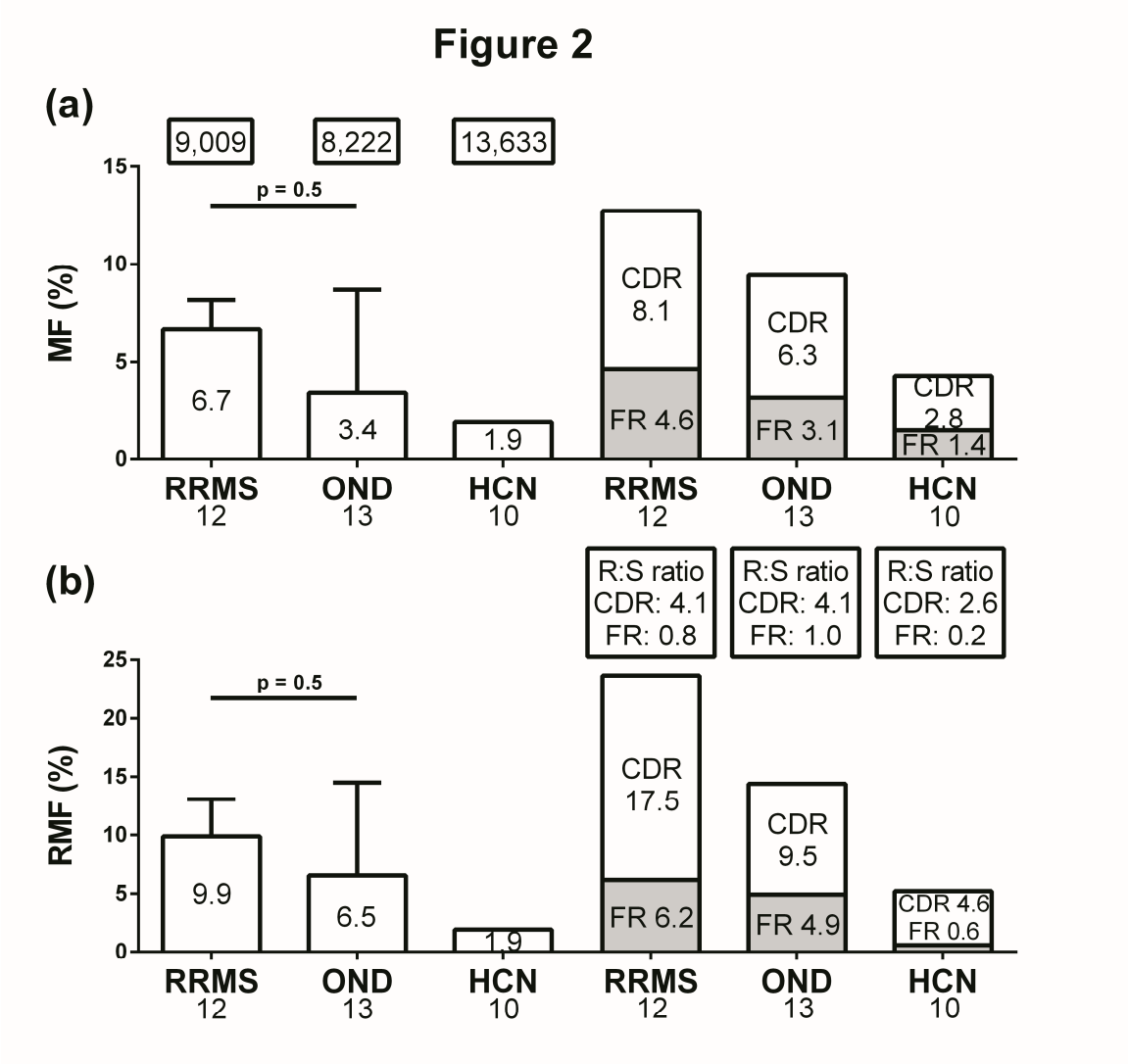


Figure 2
(a)
9,009
8,222
13,633
p = 0.5
MF (%)
15
10
5
0
6.7
3.4
1.9
CDR 8.1
FR 4.6
CDR 6.3
FR 3.1
CDR 2.8
FR 1.4
RRMS 12
OND 13
HCN 10
RRMS 12
OND 13
HCN 10
(b)
R:S ratio CDR: 4.1 FR: 0.8
R:S ratio CDR: 4.1 FR: 1.0
R:S ratio CDR: 2.6 FR: 0.2
p = 0.5
RMF (%)
25
20
15
10
5
0
9.9
6.5
1.9
CDR 17.5
FR 6.2
CDR 9.5
FR 4.9
CDR 4.6
FR 0.6
RRMS 12
OND 13
HCN 10
RRMS 12
OND 13
HCN 10

Figure 4-3.

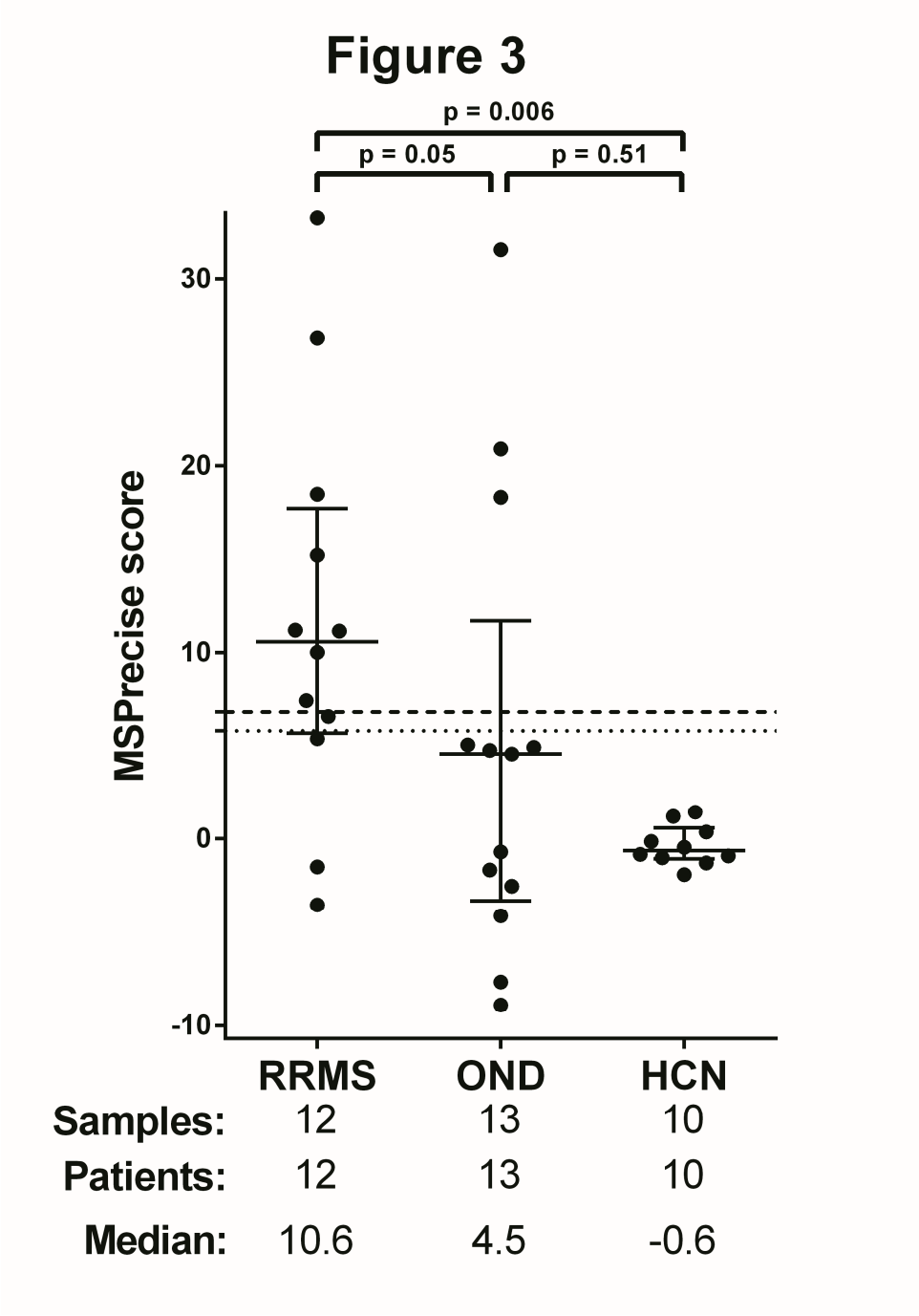


Figure 3
p = 0.006
p = 0.05
p = 0.51
MSPrecise score
30
20
10
0
-10
RRMS
OND
HCN
Samples: 12 13 10
Patients: 12 13 10
Median: 10.6 4.5 -0.6

Figure 4-4.

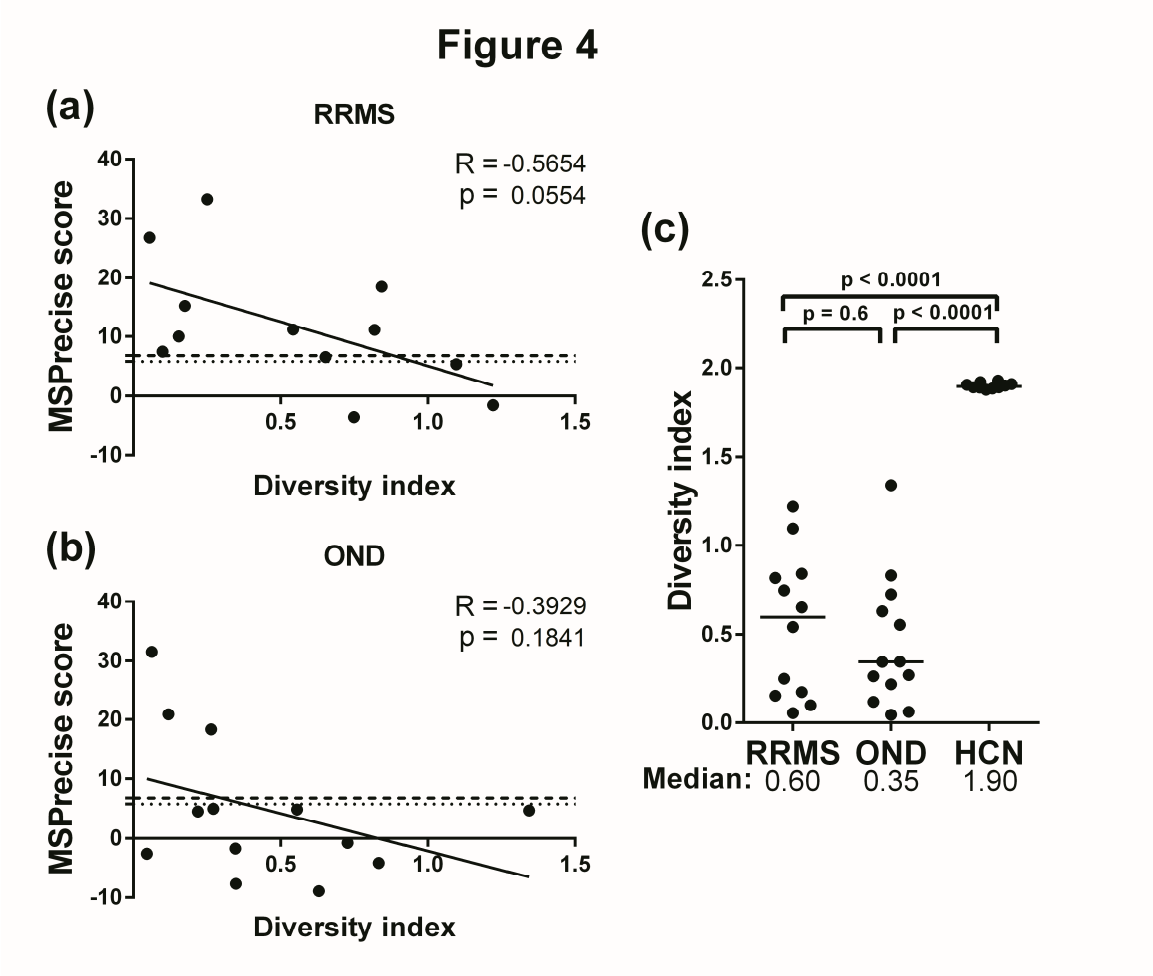


Figure 4
(a)
RRMS
R = -0.5654
p = 0.0554
MSPrecise score
Diversity index
(b)
OND
R = -0.3929
p = 0.1841
MSPrecise score
Diversity index
(c)
p < 0.0001
p = 0.6
p < 0.0001
Diversity index
RRMS
OND
HCN
Median: 0.60 0.35 1.90

Figure 4-5.

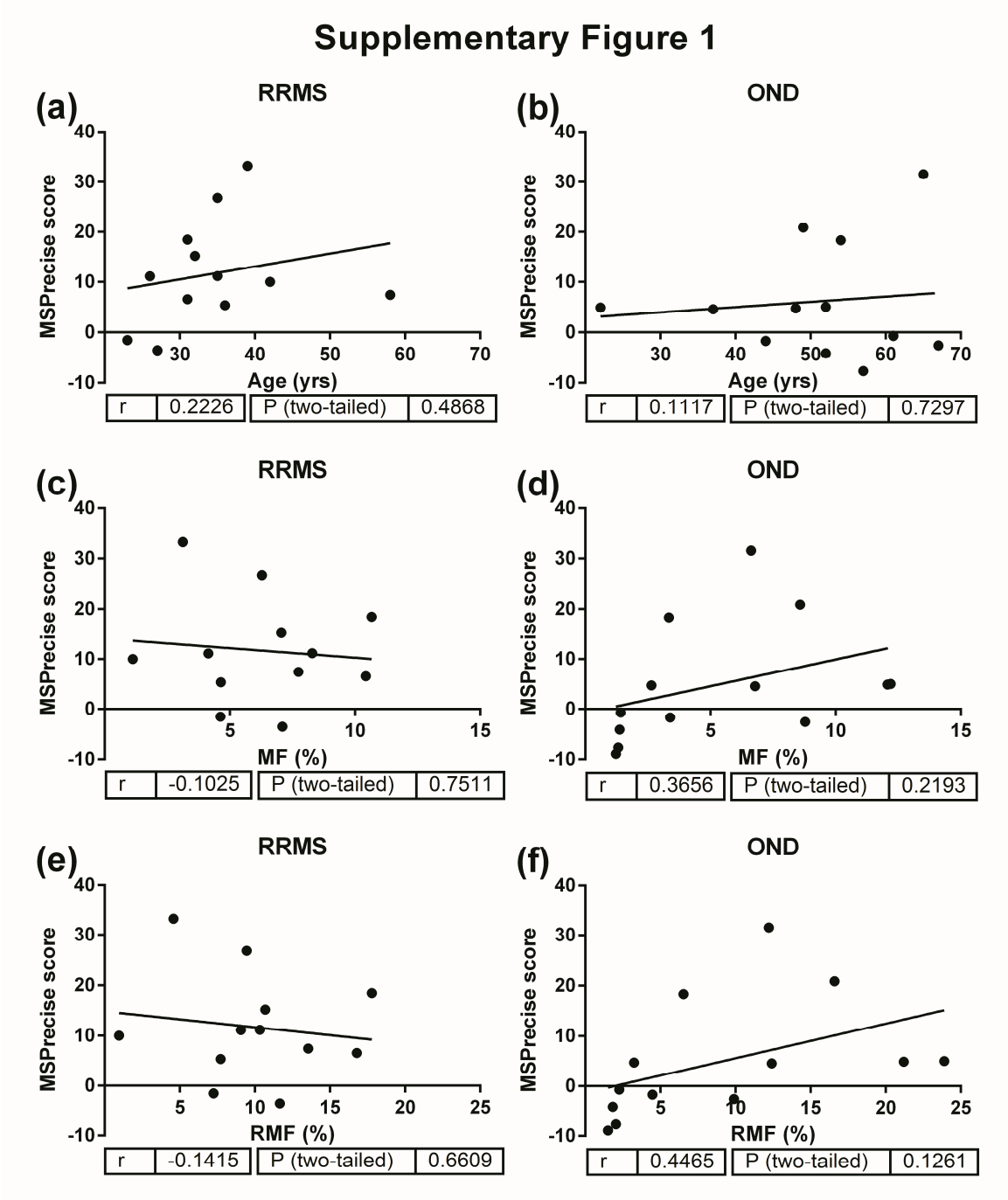

Supplementary Figure 1
(a)
RRMS
MSPrecise score
Age (yrs)
r 0.2226 P (two-tailed) 0.4868
(b)
OND
MSPrecise score
Age (yrs)
r 0.1117 P (two-tailed) 0.7297
(c)
RRMS
MSPrecise score
MF (%)
r -0.1025 P (two-tailed) 0.7511
(d)
OND
MSPrecise score
MF (%)
r 0.3656 P (two-tailed) 0.2193
(e)
RRMS
MSPrecise score
RMF (%)
r -0.1415 P (two-tailed) 0.6609
(f)
OND
MSPrecise score
RMF (%)
r 0.4465 P (two-tailed) 0.1261

Figure 4-6.

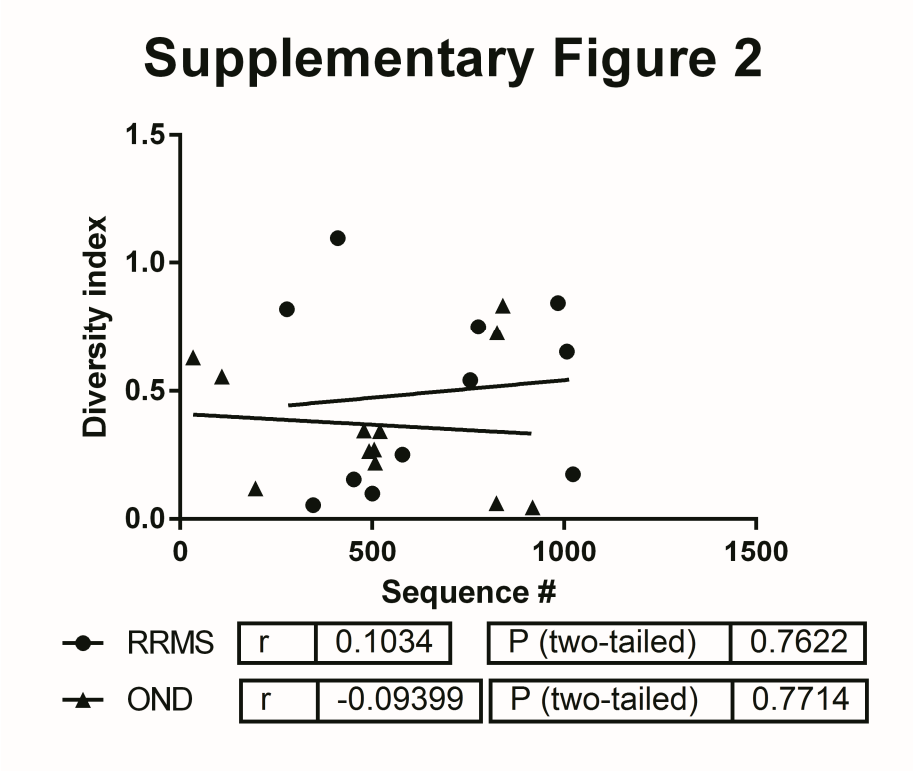
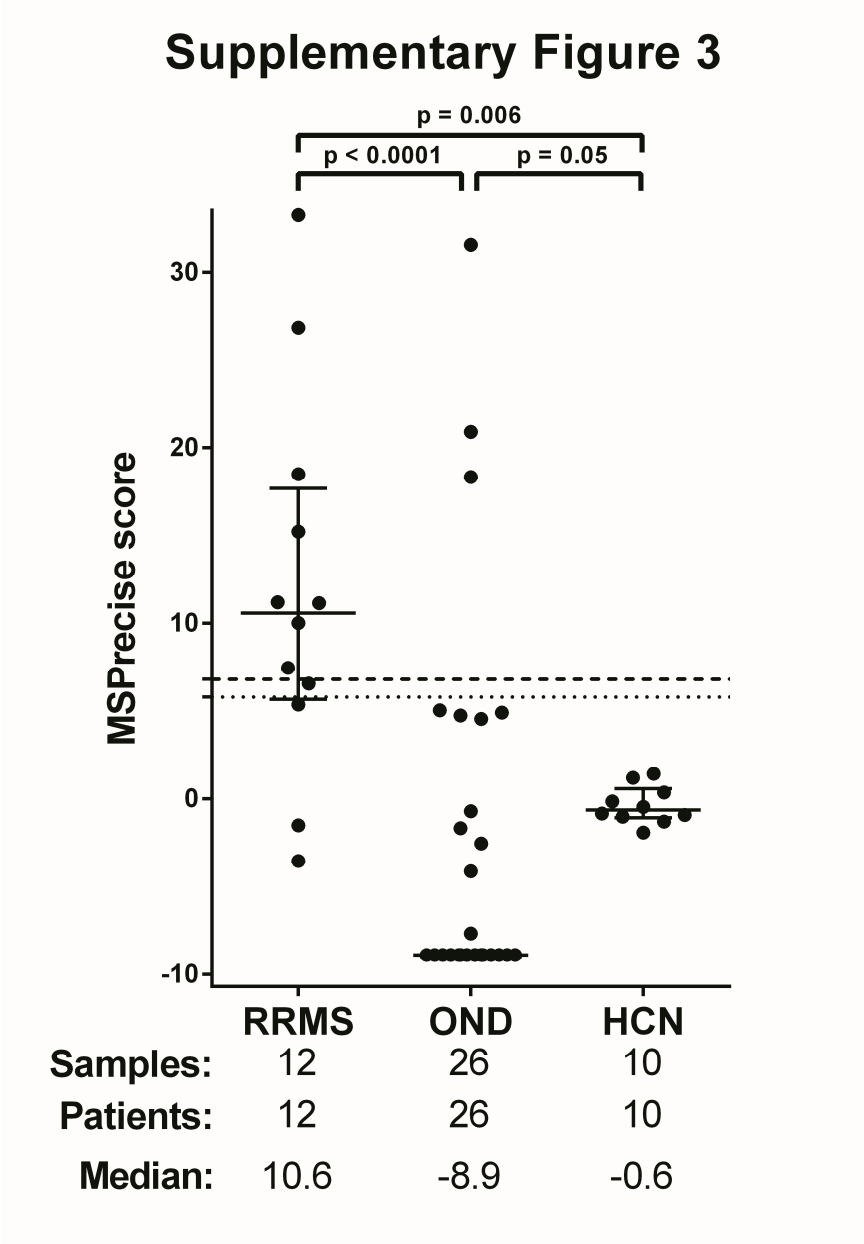

Supplementary Figure 2
Diversity index
1.5
1.0
0.5
0.0
0
500
1000
1500
Sequence #
RRMS
r
0.1034
P (two-tailed)
0.7622
OND
r
-0.09399
P (two-tailed)
0.7714

Figure 4-7.


Supplementary Figure 3
p = 0.006
p < 0.0001
p = 0.05
MSPrecise score
30
20
10
0
-10
RRMS
OND
HCN
Samples: 12 26 10
Patients: 12 26 10
Median: 10.6 -8.9 -0.6

TABLES FOR CHAPTER FOUR: RESULTS

Table 4-1. Filtering of samples by cohort.

Cohort name[a]	**RRMS**	**OND**	**HCN**[b]	**TOTAL**
Initial sample number	13	26	10	40
Samples with insufficient reads (<10 unique reads after filtering)	1[c]	13	0	14
Total analyzed	**12**	**13**	**10**	**26**

[a] Samples were grouped into patient cohorts by final diagnosis.
[b] Replicates from a single patient
[c] Patient on natalizumab at time of sampling
Abbreviations: RRMS, relapsing-remitting MS; OND, other neurological disorder; HCN, healthy control naïve peripheral B cells.

Table 4-2. (for more information, supplementary table used instead) RRMS full patient sample summary.

Patient ID	Age[a]	Gender	Diagnosis at tap	Diagnosis[b]	MS-Precise Score	Time with MS[c]	Treatment[d]	OCB status [a]
MS01	35	F	RRMS	RRMS	11.2	NR	None	NR
MS02	27	F	RRMS	RRMS	-3.54	1	steroids (7 days)	POS
MS03	39	F	RRMS	RRMS	33.25	0	None	NEG
MS04	26	F	RRMS	RRMS	11.14	76	glatiramer acetate (5 years)	NR
MS05	31	F	RRMS	RRMS	6.56	25	IFN-B1a (9 months)	POS
MS06	42	F	RRMS	RRMS	10.02	92	mycophenolic acid (7 years)	NEG
MS07	35	M	RRMS	RRMS	26.82	24	IFN-B1a (2 years)	NEG
MS08	36	F	Possible MS	RRMS	5.37	0	None	NR
MS09	31	M	RRMS	RRMS	18.5	0	None	NR
MS10	23	F	Possible MS	RRMS	-1.51	0	None	NR
MS11	32	F	RRMS	RRMS	15.22	0	None	POS
MS12	58	F	Possible MS	RRMS	7.46	0	None	NEG
MS13*	58	M	Possible MS	RRMS	-8.89	120	natalizumab (4.5 years)	NEG

[a] At time of sampling (years)
[b] Most up-to-date available
[c] At time of sampling (months)
[d] If immunomodulatory and at time of sampling
* low unique sequence read count sample (<10)
Abbreviations: NR, not reported; OCB, oligoclonal bands; RRMS, relapsing-remitting MS

Table 4-3. (for more information, supplementary table used instead) Non-RRMS full patient sample summary.

Patient ID	Age[a]	Gender	Diagnosis[b]	MSPrecise Score
OND01	37	M	OND	4.55
OND02	61	M	Dementia	-0.71
OND03	54	M	Stroke	18.34
OND04	65	F	Dementia	31.56
OND05	52	F	Headache	-4.12
OND06	48	F	Neurosarcoidosis	4.74
OND07	NR	F	Headache	-8.89
OND08	57	F	PND	-7.66
OND09	44	F	Encephalitis	-1.67
OND10	67	F	PND	-2.55
OND11	49	F	Urge incontinence	20.9
OND12	52	M	Alzheimer's	5.04
OND13	22	F	Headache	4.91
OND14*	25	F	Headache	-8.89
OND15*	40	F	Headache	-8.89
OND16*	72	F	ALS	-8.89
OND17*	56	M	CIDP	-8.89
OND18*	33	F	Suspected Glioma, possible MS	-8.89
OND19*	NR	F	Peripheral neuropathy, antiphopholipid syndrome	-8.89
OND20*	23	M	Headache, Chiari malformation	-8.89
OND21*	32	F	Sensory neuropathy	-8.89
OND22*	50	F	Hodgkin's lymphoma	-8.89
OND23*	67	F	PND	-8.89
OND24*	62	F	Dementia	-8.89
OND25*	54	F	Nondemyelinating optic neuropathy	-8.89
OND26*	58	F	PND	-8.89
HCN	NR	NR	NA	-0.65

[a] At time of sampling (years)
[b] Most up-to-date available
* low unique sequence read count sample (<10)
Abbreviations: NR, not reported; OND, other neurological disorder; PND, paraneoplastic neurologic disorder; ALS, amyothrophic lateral sclerosis; CIDP, chronic inflammatory demyelinating polyneurophathy; HCN, healthy control naïve peripheral B cells; NA, not applicable.

Table 4-4. Sequence yield per cohort.

		Redundancy 0[a]		Redundancy 1[b]			
Cohort name	N	Total unique sequences[c]	Avg. unique sequences per sample	Total unique sequences	Avg. unique sequences per sample	Avg. RM per sample	Avg. RM per sequence
RRMS	12	28,489	2,374	9,009	751	5,229	7.0
OND	13	20,201	1,554	8,222	632	3,025	4.8
HCN	10[d]	93,204	9,320	13,633	1,363	1,738	1.3
Previous MS	7	16,984	2,426	4,082	583	5,466	9.4

[a] Unique sequences with any number of reads are included in the sequence database as with the previous study (labelled "Previous MS" in the table).(Rounds et al., 2014)
[b] Filter used for this study. Unique sequences with at least two reads are included in the analysis database.
[c] After sequence filtering
[d] Replicates from a single patient
Abbreviations: RM, replacement mutation; RRMS, relapsing-remitting MS; OND, other neurological disorder; HCN, healthy control naïve peripheral B cells.

Table 4-5. AGS codon replacement mutation frequency relative to germline in RRMS and OND patients.

AGS codon	Location	RRMS RMF[a]	OND RMF	Fold higher in RRMS	p-value[b]
31B	CDR1	53.8%	38.5%	1.40	<0.001
40	FR2	16.4%	13.0%	1.26	<0.001
56	CDR2	33.6%	15.7%	2.14	<0.001
57	CDR2	21.1%	4.8%	4.42	<0.001
81	FR3	20.1%	27.0%	0.75	<0.001
89	FR3	9.3%	13.7%	0.68	<0.001

[a] Calculated relative to the total possible replacement mutations for each cohort (i.e. the number of reads that have a specific numbered codon in the germline)
[b] Calculated by Chi-squared test
Abbreviations: RMF, replacement mutation frequency; RRMS, relapsing-remitting MS; OND, other neurological disorder.

CHAPTER FIVE

RESULTS

AIM II: AGS is a unique feature of disease

Overview and rationale

Previous work established that the MSPrecise tool (AGS scoring in a clinical setting on CSF cell pellets) had good performance on a cohort of 39 patients. As a result, the next testing benchmark for MSPrecise was a validation clinical trial. Previous work had established that sequence amplification during PCR was a key area for optimization to reduce repertoire skewing from over-amplified templates. For this purpose, the PCR protocol was switched to a single targeted VH4 PCR step prior to barcode attachment. Additionally, changes were made to the sequence analysis method to further reduce the impact of over amplification. These changes revolved around the definition of a "unique" read (see Methods; Table 2-1), a key feature since unique reads that only have one corresponding raw sequence template are removed from the database. In addition, to ensure the most accurate patient cohort determination possible, patient diagnosis was determined by a panel of 3 independent adjudicators to ensure that MSPrecise accuracy would be evaluated against patients clearly with or without MS.

VALIDATION TRIAL FOR A GENETICS-BASED ADD-ON DIAGNOSTIC TEST FOR MULTIPLE SCLEROSIS.

The following study is being submitted for publication. Rounds WH, Wilks TB, 2nd, Corboy JR, Ratchford JN, Murray RS, Gudesblatt M, Bigwood DW, Eastman EM, Greenberg BM, Cowell LG, Monson NL, *Validation trial for a genetics-based add-on diagnostic test for multiple sclerosis.*, 2016.

Introduction

Multiple sclerosis (MS) is a complex disease of the central nervous system (CNS) with pathology related to both autoimmunity and failure of repair (Frohman et al., 2006a; Frischer et al., 2009). Early diagnosis continues to be a primary goal (McDonald et al., 2001), as MS patients exhibit better prognoses when treated earlier (Frohman et al., 2006b). Currently, radiological testing and patient history are the core tools for clinicians seeking to confirm an MS diagnosis (Milo and Miller, 2014). Detection of oligoclonal bands (OCB) in the cerebrospinal fluid (CSF) is also used, but the OCB test is primarily limited in its specificity when trying to distinguish between MS patients and those with other neuro-inflammatory diseases (roughly 61% specificity) (Reske et al., 2005; Tintore et al., 2008; Petzold, 2013).

OCB indicate that antibodies are being produced by terminally differentiated B cells called plasmablasts or plasma cells in the CNS. In 2007, no therapies targeted this B cell subtype. Since then, three B cell depleting therapies (BCDT), Rituximab,

Ocrelizumab and Ofatumumab, which target the precursors of plasmablasts and plasma cells as well as less differentiated B cell subtypes, demonstrated tremendous efficacy in the treatment of relapsing remitting MS patients (Hauser et al., 2008; Kappos et al., 2011; Sorensen et al., 2014). While BCDT does not affect total levels of CSF IgG or reduce OCB count during the initial response to BCDT (Cross et al., 2006), it does reduce the number of B and T cells in the CSF, as well as the concentration of certain chemoattractants for B and T cells (Cross et al., 2006; Piccio et al., 2010). This suggests that B cell involvement in MS is through their function as antigen-presenting cells.

The function of a B cell from early development to effector status is largely driven by the antibody the B cell produces (Meffre et al., 2000; Gauld et al., 2002). Thus, our laboratory has focused on the antibody genetics of antigen-experienced B cells from MS patients as a means to understand whether they have engaged proper selection mechanisms to achieve effector status. For example, previous work by our laboratory (Monson et al., 2005b; Harp et al., 2007) and others (Owens et al., 1998; Qin et al., 1998; Baranzini et al., 1999; Colombo et al., 2000; Owens et al., 2003; Owens et al., 2007) demonstrated that B cells in the CSF of MS patients accumulate more mutations and are enriched for variable heavy chain family 4 (VH4) family genes. These B cells have also undergone clonal expansion in the CNS (Corcione et al., 2004; Serafini et al., 2004; Monson et al., 2005b). More recently, we hypothesized that B cells in the CSF of MS patients express a distinct pattern of somatic hypermutation (SHM) characteristic of their putative exposure to neuro-antigens. Indeed, in these early studies, we reported that 6 codons in the immunoglobulin VH4 gene segments had significantly elevated rates of replacement mutation frequency in CSF-derived VH4+ B cells isolated from the CSF of

MS patients. In fact, when applied to patient cohorts, the prevalence of SHM accumulation at these 6 codons was 91% accurate in identifying patients with relapsing-remitting multiple sclerosis (RRMS) and clinically isolated syndromes (CIS) who would subsequently convert to RRMS (Cameron et al., 2009; Rounds et al., 2014).

We have since converted this technology to next generation sequencing (NGS) and renamed the test MSPrecise®, which uses 5 out of the 6 initial codons after demonstrating that the sixth codon has lower replacement mutation frequency (RMF) by NGS in the RRMS cohort than any of the other 5 (Rounds et al., 2015). The study presented here is an evaluation of MSPrecise® in a larger patient cohort and of the underlying antibody genetics of CSF-derived B cells from patients with either RRMS or other neurological disorder (OND). Patient samples were obtained from thirteen different clinical centers and categorized as definite RRMS, definite OND, or unclear by three independent adjudicators based on the patients' existing clinical features, reported MRI data and history. Out of the 146 patients who yielded sufficient sequence data, the 3 adjudicators reached a consensus on the RRMS or OND diagnosis for 76 patients (Figure 5-1). Patients with consensus of diagnosis for 2 out of 3 adjudicators were used only for antibody genetics analysis. We hypothesize that MSPrecise® scores are higher in the RRMS cohort compared to the OND cohort and that this clinical test is an effective add-on diagnostic tool for identifying RRMS patients.

Results

Cohort composition

All patients consented to this study under institutional review board guidelines. The independent adjudicators evaluated the 146 patients who participated in this study and categorized them as “RRMS”, “OND” or “unclear”. All three adjudicators agreed on a diagnosis of RRMS for 41 patients ($RRMS_{3/3}$ – 3/3 votes), and a diagnosis of OND for 22 patients ($OND_{3/3}$ – 3/3 votes) which were used for MSPrecise® accuracy evaluation (Figure 5-1). An additional 16 RRMS patients and 18 OND patients were identified by 2 of the 3 adjudicators ($RRMS_{2/3}$ and $OND_{2/3}$). Quality analysis of the sequence data from these 57 RRMS patients and 40 OND patients led us to exclude 5 RRMS and 13 OND samples from the antibody genetics analysis due to a low number of sequence reads (Figure 5-1). Thus, a total of 52 RRMS and 27 OND patient samples were included in the antibody genetics feature analysis summarized in Figures 5-2 through 5-4.

RRMS CSF-infiltrating B cells undergo increased affinity maturation compared to OND

Since this is the largest VH4 antibody gene library ever reported for B cells from the CSF of RRMS patients, we analyzed the repertoires to identify features of antibody genetics that may be distinct in RRMS patients compared to OND patients. Overall, the RRMS cohort contained more VH4+ sequences per patient than the OND cohort (Figure 5-2A; median 67 in RRMS vs 31 in OND, $p=0.006$). This was largely due to the observation that 13/40 of the OND samples had very low sequence counts (defined as no more than 8 sequences per sample, or no more than 24 sequences per sample with high amplification bias). When the samples with low sequence counts were removed from the analysis, sequence counts in the RRMS and OND cohorts used for all subsequent antibody genetics analysis are no longer statistically different (Figure 5-2B). We also

observed a higher frequency of clonally related B cells in the CSF of RRMS patients compared to the OND cohort, with a median of 52% unique sequences sharing a matching CDR3 compared to a median of 38% in the OND cohort (Figure 5-2C; p = 0.026). Furthermore, the number of distinct clones in the RRMS samples was significantly elevated compared to OND (Figure 5-2D; median 15.5 in RRMS vs 8 in OND, p=0.04). This observation recapitulates what we (Rounds et al., 2015) and others (Harp et al., 2007; Owens et al., 2007; Bennett et al., 2008) previously observed using single cell Sanger sequencing.

Despite lower sequence frequency and decreased clonal expansion in the OND cohort compared to the RRMS cohort, the productive VH4 antibody genes expressed by B cells in the CSF of RRMS and OND patients did not significantly differ in mutation frequency (MF) (Figure 5-3A; 5.5 RRMS to 4.4 OND, p=0.11). Both the RRMS and OND cohorts had proper targeting of mutations since the R:S ratios in the CDRs (medians 4.5 for RRMS and 3.7 for OND) were higher than in the FRs (medians 1.1 for RRMS and 1.1 for OND). The overall RMF in the RRMS cohort was not significantly higher in comparison to the OND cohort (Figure 5-3B; 10.1 RRMS to 7.0 OND, p=0.08). Further analysis of this difference in RMFs (Figures 5-3C) revealed a greater proportion of sequences with 2 or more replacement mutations in the RRMS cohort compared to the OND cohort (medians 94% to 85%, p = 0.03).

VH4 gene usage within the RRMS repertoires are distinct from the OND repertoires

Further analysis of the distribution of VH4 gene segments was performed to determine whether particular over and underrepresented genes were enriched in either the

RRMS or OND cohorts (Figure 5-4A). Indeed, the distribution of VH4 gene segments within the RRMS and OND cohorts were very distinct from each other (Chi-squared value = 605 [only value of this Chi-square is in comparison to the JH family to show representation of divergence versus similarity]). In particular, the VH4 gene segments utilized by the RRMS cohort was different from a uniform distribution such that VH4-39 (percentage deviation = 69%) was used more frequently than expected in the RRMS cohort. This finding also confirmed what we (Rounds et al., 2015) and others (Owens et al., 2007) have previously observed regarding VH4 gene segment distribution in RRMS patients. The VH4 gene segments utilized by the OND cohort was also different from a uniform distribution, with a pronounced overrepresentation of VH4-59 (percentage deviation = 95%). In contrast, both the RRMS and OND cohorts displayed similar JH gene distribution (Chi-squared value = 209), with the largest portion of sequences aligned to JH4 (36-42%) and about 37-47% of all sequences utilizing the distal JH5 and JH6 genes (Figure 5-4B).

MSPrecise® identifies RRMS patients with high sensitivity and specificity

The performance of the MSPrecise® score was evaluated using the $RRMS_{3/3}$ and $OND_{3/3}$ cohorts in which 3 of the 3 adjudicators agreed on diagnosis in order to avoid any skewing of the data due to incorrect cohort assignment. We calculated MSPrecise® scores for all 63 patient CSF samples from the $RRMS_{3/3}$ and $OND_{3/3}$ cohorts (Figure 5-5). Previously, we established a NGS MSPrecise® threshold of 5.8, such that samples with MSPrecise® scores above 5.8 were predicted to be RRMS (Rounds et al., 2015). We also determined in these earlier studies that samples with low sequence count should be

identified as OND (automatic score assignment of -8.889). In this study, the median MSPrecise® score of the $RRMS_{3/3}$ cohort was statistically higher than the $OND_{3/3}$ cohort (9.7 vs -8.4; p=0.0002). Of the 41 $RRMS_{3/3}$ samples, 32 (78%) had MSPrecise® scores above this threshold. Of the 22 $OND_{3/3}$ samples, 17 (77%) had MSPrecise® scores below this threshold. The remaining 5 $OND_{3/3}$ samples, which had high MSPrecise® scores, were adjudicated as Lyme disease (n=1), pseudo-tumor (n=1), headache (n=1) and unknown diagnosis (n=2). Of the 10 $OND_{3/3}$ patients that were adjudicated as headache, 9 (90%) had MSPrecise® scores below the 5.8 threshold.

Although our previous study noted that higher overall RMF values can influence the MSPrecise® score (Rounds et al., 2015), no correlations were observed in either cohort between MSPrecise® score and RMF (data not shown). This previous study also associated high diversity of VH4 gene usage with low MSPrecise® scores. However, the diversity index (DI) for this study was not an issue due to the sequence normalization by gene and RM pattern we use here (median DI for $RRMS_{3/3}$ 1.2; 0.6 for previous study). In addition, in this study we determined that the average and median incidence of RMs per sequence at each of the 5 codons used to calculate MSPrecise® score (31B, 40, 56, 57, 81) was higher in $RRMS_{3/3}$ samples compared to $OND_{3/3}$ for each codon position (data not shown).

Based on these criteria for MSPrecise®, the validation study presented here demonstrated a sensitivity of 78% for the $RRMS_{3/3}$ cohort, a specificity of 77% for the $OND_{3/3}$ cohort and an overall accuracy of 78% (Figure 5-5). Interestingly, 40 of the 41 $RRMS_{3/3}$ samples and 21 of the 22 $OND_{3/3}$ samples also had OCB test results available. When both MSPrecise® and OCB test results matched (29 $RRMS_{3/3}$ samples and 15

$OND_{3/3}$ samples), the sensitivity increases to 93% and the specificity to 93% for an overall accuracy of 93%. We also noted that within the $OND_{3/3}$ subset of patients who had a headache diagnosis (n=10), the specificity of MSPrecise® was increased to 90% and none of the headache patients had false positive test results for both MSPrecise® and OCB simultaneously.

Discussion

This study is the largest genetic analysis of the VH4+ B cells producing antibodies from the CSF of RRMS patients in comparison to ONDs. Our first observation was that there was a greater proportion of OND patients that have little to no sequence recovery from putative CSF-derived B cells compared to the RRMS cohort (13/40=33% OND compared to 5/57=9% RRMS). This observation is not unexpected, as RRMS is characterized by a weakening of the blood-brain barrier that allows for lymphocyte infiltration in the CSF (Minagar and Alexander, 2003; Holman et al., 2011). In contrast, patients in the headache subgroup had low sequence recovery in most samples (6/10), further supporting the concept that blood-brain barrier weakening is not necessarily associated with headache diagnosis. Indeed, others have reported in case studies that the frequency of lymphocytes in the CSF of headache patients is quite variable (Filina et al., 2013).

Another striking feature of the RRMS cohort in this study is that it exhibits increased genetic features of clonal expansion (Figure 5-2) and selection (Figure 5-3) compared to the OND cohort. Even when the higher number of samples with low sequence counts in the OND cohort are removed, the RRMS repertoire displays both

increased B cell CDR3 diversity as well as greater numbers of B cells with shared CDR3s (Figure 5-2). Additionally this expanded RRMS CSF B cell repertoire has more sequences with a memory B cell phenotype compared to OND (Figure 5-3).

We and others have reported these same features in RRMS patients, but comparisons to OND populations has not been robust. For example, prior to this study, the ratio of RRMS repertoires to OND repertoires reported in the literature was 2.4:1 (107 RRMS to 45 OND) (Qin et al., 1998; Colombo et al., 2000; Colombo et al., 2003; Owens et al., 2003; Ritchie et al., 2004; Monson et al., 2005b; Harp et al., 2007; Owens et al., 2007; Winges et al., 2007; von Budingen et al., 2008; Cameron et al., 2009; von Budingen et al., 2010; von Budingen et al., 2012; Ligocki et al., 2013; Palanichamy et al., 2014; Rounds et al., 2014). We would reason that our interpretation of these findings, that B cells in the CSF of RRMS patients are more abundant compared to B cells in the CSF of ONDs, is substantiated since the ratio of RRMS repertoires to OND repertoires in this study is 1.4:1 (57 RRMS to 40 OND).

Although the criteria for RRMS diagnosis have been clearly established (McDonald et al., 2001) and updated (Milo and Miller, 2014), in practice, distinguishing between RRMS and OND patients can still prove to be a challenge. This is evident in our study since the 3 independent adjudicators only reached consensus for 77/146 (52.7%) of the patients we enrolled (45 RRMS, 31 OND and 1 CIS; prior to any sample filtering by diagnosis or quality). This data emphasizes the need for supportive biomarkers to diagnose CNS diseases, including RRMS. Thus, our laboratory endeavored to harness antibody genetics as a means to support RRMS diagnosis, resulting in the development of MSPrecise® (Rounds et al., 2015).

The NGS platform presented some unique challenges, and required us to establish several core parameters for using MSPrecise® with NGS data as opposed to the single cell Sanger sequencing platform from which MSPrecise® was first developed (Cameron et al., 2009). These include a score threshold of 5.8, a requirement for identical sequences to be present at least twice in a patient repertoire to be included, and a 5 codon scoring system. In this study, we further refined our sequence processing by defining a unique repertoire sequence as defined by its VH4 gene, JH family, CDR3 and RM pattern. This helped us further reduce amplification bias through two separate mechanisms: over amplified sequences are more likely to be represented by more than one mismatching sequence due to background, and under sequenced real DNA reads are more likely to be included since the grouping by RM pattern will help them rise above the threshold of two matching reads. As a result, DI is increased across the RRMS and OND cohorts in this study relative to the previous study (Rounds et al., 2015). This change was most noticeable with the OND cohort, as we would expect since lower starting cell count will increase amplification bias differences over the course of PCR (previous OND DI median = 0.35; current for OND = 0.89). Similarly, the genetic features characteristic with increased response to antigen exhibited by the RRMS repertoire were not as readily distinguishable from the OND cohort in our previous study, due to the greater range of MF and RMF values exhibited by the OND cohort despite clearly lower medians.

The observed accuracy of OCB in the $RRMS_{3/3}$ and $OND_{3/3}$ cohorts with a reported OCB test was 85%, which is higher than its established low diagnostic specificity when comparing test performance for MS versus other neuro-inflammatory diseases (about 61%) (Reske et al., 2005; Tintore et al., 2008; Petzold, 2013). This is

most likely due to its use by the adjudicators for diagnosis. For instance, samples with RRMS clinical presentation, but with negative OCB are less likely to be classified as definitively RRMS by the independent adjudicators. As a result, we included our analysis of OCB primarily to evaluate the combined accuracy of OCB and MSPrecise®. Of the 40 $RRMS_{3/3}$ patients with a reported OCB, only 2 were identified by both tests as OND, and 11 were identified incorrectly by only one test (4 OCB negative, 7 MSPrecise® negative). Of the 21 $OND_{3/3}$ patients with a reported OCB, only 1 was identified by both tests as RRMS, and 6 were identified incorrectly by only one test (2 OCB positive, 4 MSPrecise® positive).

Acknowledgements

The authors wish to thank the patients who provided samples for this study. We thank Drs. Alan Martin, Mark Tullman, William Hu, Jacob Sloan, Rip Kinkel, Donna Graves, John Huddlestone, Carolina Ionete, Peter Riskind, Mike Racke, Jerome Graber, Jeff Kaplan, Rebecca Romero and David Brandes for coordinating sample collection at the 13 sites. We also thank Dr. Lee Szkotnicki and SeqWright Genomic Services (a GE company, Houston, TX) for performing sequencing to generate the NGS database.

FIGURE LEGENDS FOR CHAPTER FIVE: RESULTS

Figure 5-1. Patient flow diagram for the study. Patient sample filtering is shown for all patient samples sequenced in this study (duplicates of patient samples are only counted once). 3/3 and 2/3 adjudicated identifies how many independent adjudicators agreed with a diagnosis. Primary Progressive MS (PPMS), possible MS and Neuromyelitis Optica (NMO) were all excluded from the OND cohorts. Excluded samples are indicated by dashed-margin boxes and final sample counts used for either MSPrecise® or antibody genetics analyses are marked in bold.

Figure 5-2. RRMS sequence repertoires are more clonally enriched compared to OND. RRMS and OND samples identified by 2 or more adjudicators were included and low sequence samples were excluded. (A,B) Sequences per patient sample, each represented by one point. (B) Samples with low sequence count removed. (C) Clones were identified as sequences with matching CDR3 nucleotide reads within one sample. Percentages represent the fraction of sequences that are clonally related within each sample. (D) Total count of unique CDR3 nucleotide reads within one sample. Statistical significance of the difference between cohorts was calculated by Mann Whitney test.

Figure 5-3. RRMS sequence repertoires display more affinity maturation compared to OND. RRMS and OND samples identified by 2 or more adjudicators were included and low sequence samples were excluded. (A) Average mutation frequency (MF) for each patient was calculated as total mutated nucleotides divided by total sequence lengths. (B) Frequency and region targeting of codon replacement mutations (RMF).

RMF and replacement to silent mutation ratios (R:S) for CDR and FR regions were calculated independently by region for each sample. R:S ratios are shown as cohort medians. (C) Percentage of sequences per patient with 0, 1 or 2+ replacement mutations. Statistical significance of the difference between cohorts was calculated by Mann Whitney test.

Figure 5-4. RRMS and OND CSF B cells show discordance in VH4 gene distribution but similar JH gene distribution. RRMS and OND samples identified by 2 or more adjudicators were included and low sequence samples were excluded. Total cohort sequence counts are indicated below respective pie charts. VH4 (A) and JH (B) gene calls were obtained by IgBlast alignment through VDJServer (see methods). Chi-squared analysis values between cohort gene distributions are shown above the bars. Gene frequencies are shown in the table.

Figure 5-5. MSPrecise® scores distinguish between RRMS and OND patient cohorts. RRMS and OND samples identified by all 3 adjudicators were included and low sequence samples were assigned a minimum score of -8.89. The **MSPrecise®** cut-off point of 5.8 is indicated by a dotted line. The headache subgroup of the OND cohort was also included for comparison to the RRMS cohort. Statistical significance of the difference between cohorts was calculated by Mann Whitney test.

FIGURES FOR CHAPTER FIVE: RESULTS

Figure 5-1.

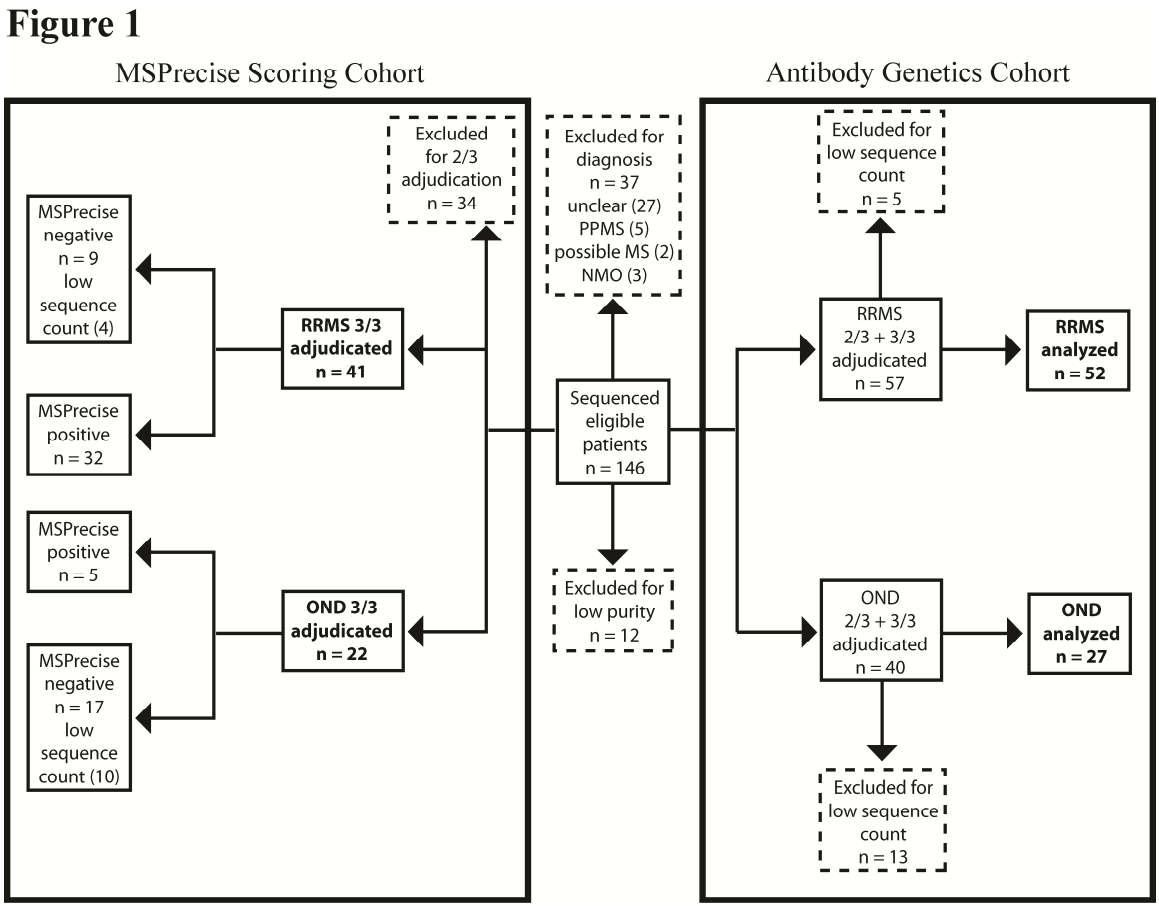


Figure 5-2.

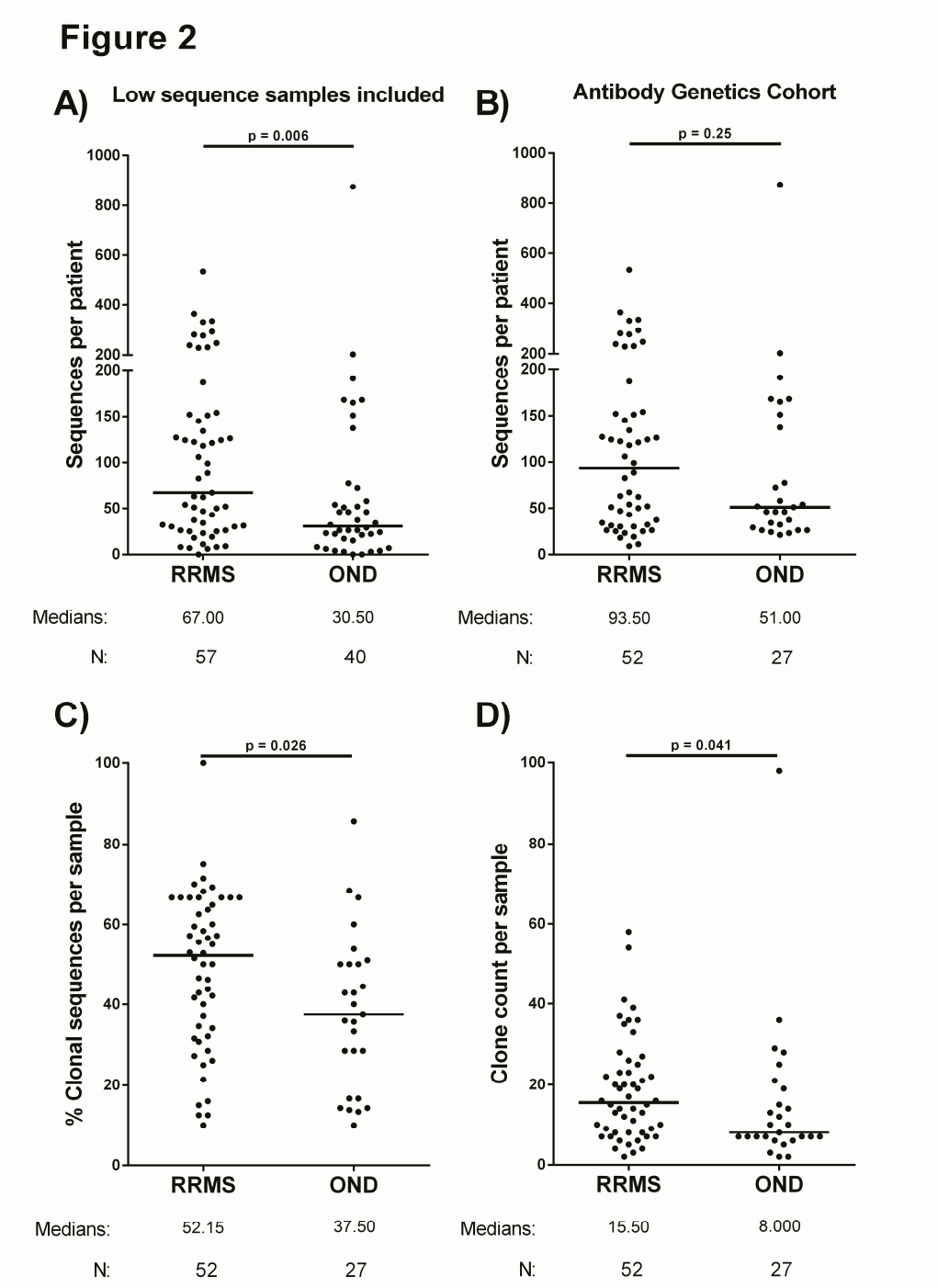


Figure 2
A)
Low sequence samples included
p = 0.006
Sequences per patient
RRMS
OND
Medians: 67.00 30.50
N: 57 40
B)
Antibody Genetics Cohort
p = 0.25
Sequences per patient
RRMS
OND
Medians: 93.50 51.00
N: 52 27
C)
p = 0.026
% Clonal sequences per sample
RRMS
OND
Medians: 52.15 37.50
N: 52 27
D)
p = 0.041
Clone count per sample
RRMS
OND
Medians: 15.50 8.000
N: 52 27

Figure 5-3.

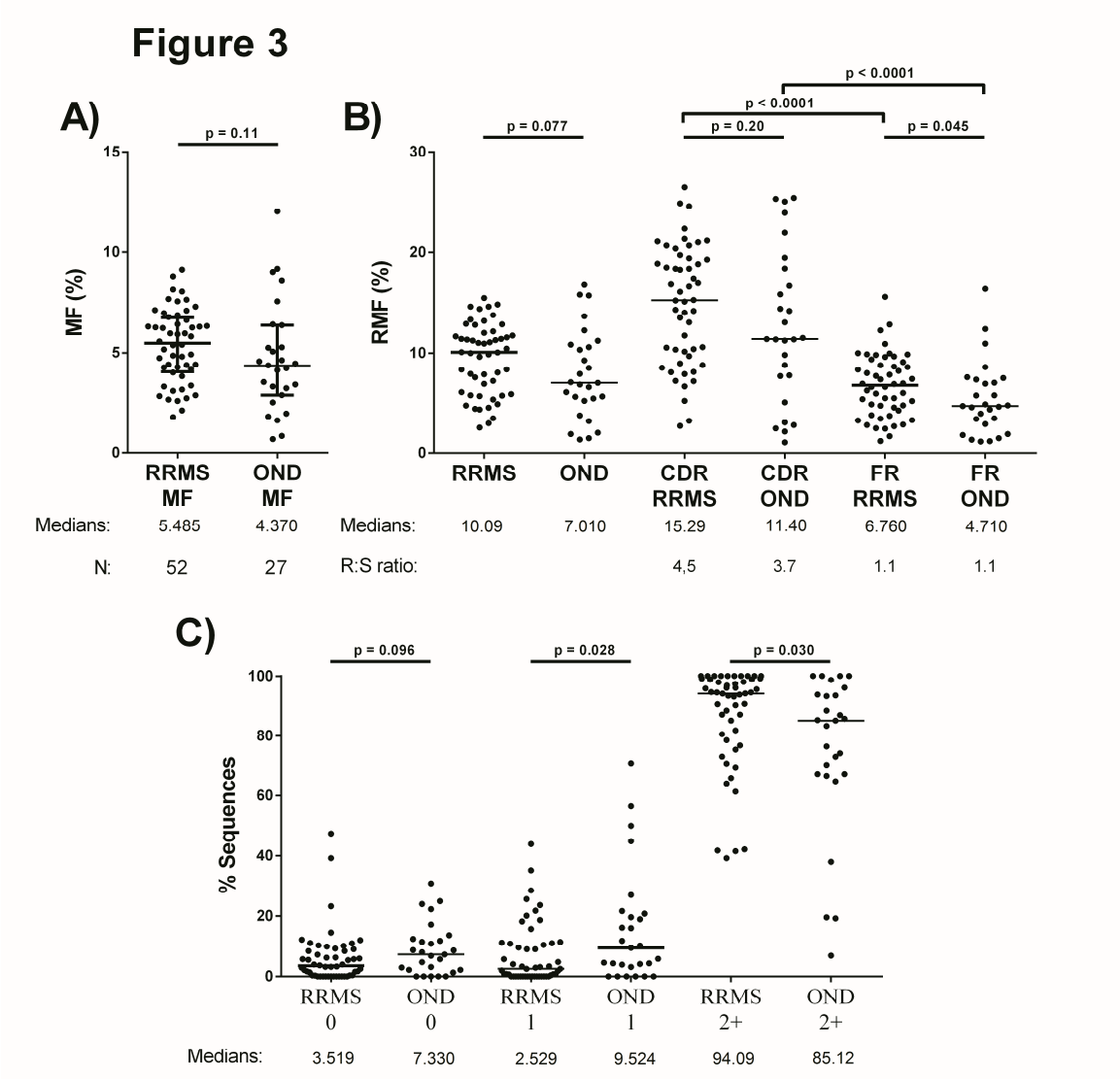


Figure 3
A)
p = 0.11
MF (%)
15
10
5
0
RRMS MF
OND MF
Medians: 5.485 4.370
N: 52 27
B)
p = 0.077
p < 0.0001
p = 0.20
p < 0.0001
p = 0.045
RMF (%)
30
20
10
0
RRMS
OND
CDR RRMS
CDR OND
FR RRMS
FR OND
Medians: 10.09 7.010 15.29 11.40 6.760 4.710
R:S ratio: 4,5 3.7 1.1 1.1
C)
p = 0.096
p = 0.028
p = 0.030
% Sequences
100
80
60
40
20
0
RRMS 0
OND 0
RRMS 1
OND 1
RRMS 2+
OND 2+
Medians: 3.519 7.330 2.529 9.524 94.09 85.12

Figure 5-4.

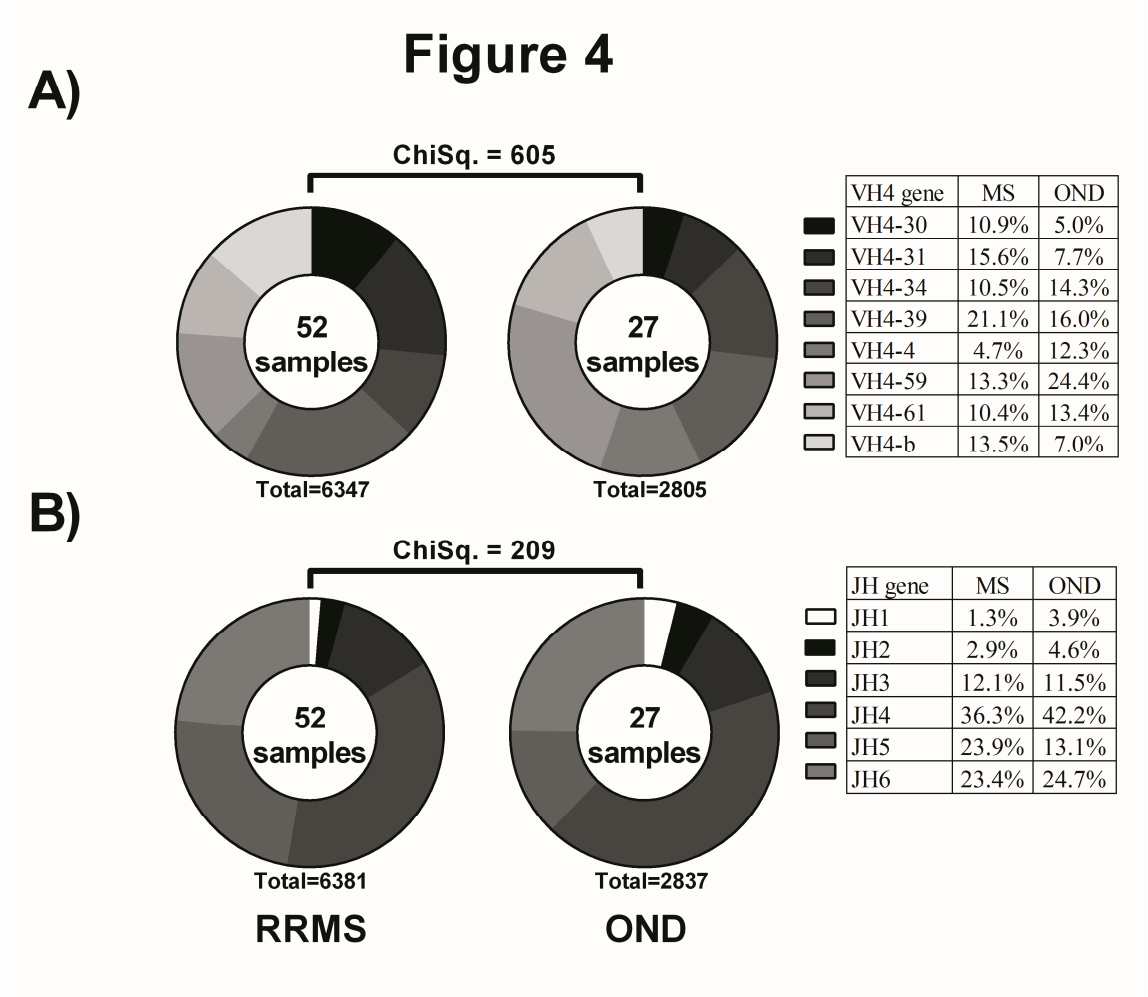
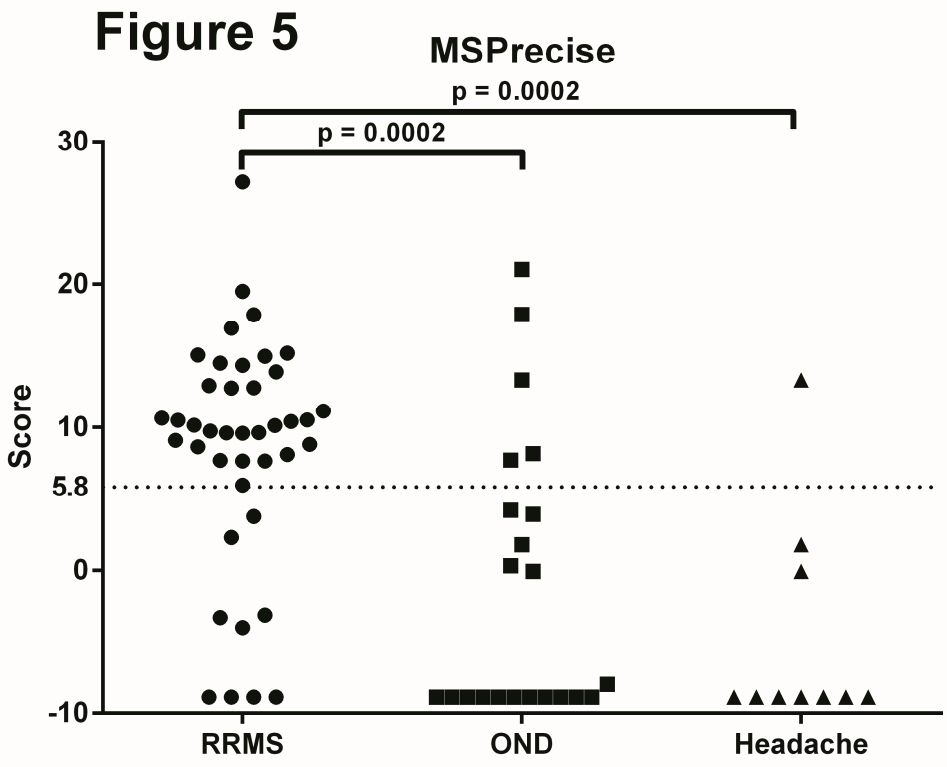


Figure 4
A)
ChiSq. = 605
52 samples
27 samples
Total=6347
Total=2805
VH4 gene | MS | OND
VH4-30 | 10.9% | 5.0%
VH4-31 | 15.6% | 7.7%
VH4-34 | 10.5% | 14.3%
VH4-39 | 21.1% | 16.0%
VH4-4 | 4.7% | 12.3%
VH4-59 | 13.3% | 24.4%
VH4-61 | 10.4% | 13.4%
VH4-b | 13.5% | 7.0%
B)
ChiSq. = 209
52 samples
27 samples
Total=6381
Total=2837
JH gene | MS | OND
JH1 | 1.3% | 3.9%
JH2 | 2.9% | 4.6%
JH3 | 12.1% | 11.5%
JH4 | 36.3% | 42.2%
JH5 | 23.9% | 13.1%
JH6 | 23.4% | 24.7%
RRMS
OND

Figure 5-5.



RRMS	Sensitivity	Specificity	Accuracy	Sample count	
versus	**MSPrecise test alone**			41	RRMS
OND	78%	77%	78%	22	OND
Headache	78%	90%	80%	10	Headache
	OCB test alone			40	RRMS
OND	85%	86%	85%	21	OND
Headache	85%	89%	86%	9	Headache
	MSPrecise and OCB match			29	RRMS
OND	93%	93%	93%	15	OND
Headache	93%	100%	94%	7	Headache

TABLES FOR CHAPTER FIVE: RESULTS

Table 5-1. RRMS sample list for MSPrecise®

Diagnosis	Sequences	MSP	OCB		Diagnosis	Sequences	MSP	OCB
RRMS 01	121	-3.33	Pos		RRMS 22	31	13.90	Pos
RRMS 02	99	10.41	Pos		RRMS 23	19	3.74	Neg
RRMS 03	276	12.78	Pos		RRMS 24	51	10.66	Pos
RRMS 04	50	12.77	Pos		RRMS 25	9	9.08	Pos
RRMS 05	364	9.62	Pos		RRMS 26	152	8.63	Pos
RRMS 06	63	2.27	Pos		RRMS 27	331	9.73	Pos
RRMS 07	335	7.68	NA		RRMS 28*	8	-8.89	Pos
RRMS 08	238	5.93	Pos		RRMS 29*	0	-8.89	Pos
RRMS 09	30	12.94	Pos		RRMS 30	126	16.94	Neg
RRMS 10	151	10.13	Pos		RRMS 31	54	17.88	Pos
RRMS 11	122	14.98	Pos		RRMS 32	292	10.15	Neg
RRMS 12	187	14.35	Neg		RRMS 33	52	9.58	Pos
RRMS 13	145	11.15	Pos		RRMS 34*	7	-8.89	Pos
RRMS 14	127	8.80	Pos		RRMS 35	280	9.61	Neg
RRMS 15	82	-3.17	Neg		RRMS 36	106	7.64	Pos
RRMS 16	11	27.19	Pos		RRMS 37	228	10.49	Pos
RRMS 17	62	15.21	Pos		RRMS 38	18	7.62	Pos
RRMS 18*	8	-8.89	Pos		RRMS 39	230	-4.02	Pos
RRMS 19	118	10.51	Pos		RRMS 40	23	8.09	Pos
RRMS 20	30	15.06	Pos		RRMS 41	154	19.50	Pos
RRMS 21	25	14.50	Pos					

Abbreviations: Pos, positive; Neg, negative; NA, not available.
*samples with low sequence count automatically scored as -8.89

Table 5-2. OND sample list for MSPrecise®

Diagnosis	Sequences	MSP	OCB	Additional information
OND 01	137	7.68	Neg	Lyme disease 3/3
OND 02	0	-8.89	Neg	Neurosarcoidosis 3/3
OND 03*	0	-8.89	Neg	Headache 3/3
OND 04*	875	3.90	Neg	
OND 05	168	0.32	Neg	Lupus 3/3
OND 06*	3	-8.89	Neg	Palsy 2/3
OND 07*	26	1.77	Neg	Headache 3/3
OND 08	24	17.89	Neg	Pseudotumor 3/3
OND 09	17	-8.89	Neg	Headache 2/3
OND 10*	26	-8.89	Neg	Headache 2/3
OND 11	37	8.14	Pos	
OND 12	72	-0.07	Neg	Headache 2/3
OND 13	7	-8.89	Neg	Headache 3/3
OND 14*	15	-8.89	Neg	Headache 2/3
OND 15*	3	-8.89	NA	Headache 2/3
OND 16	32	21.03	Neg	
OND 17	58	-7.98	Neg	Hepatitis 2/3
OND 18	8	-8.89	Neg	
OND 19	6	-8.89	Pos	Headache 3/3
OND 20	22	-8.89	Pos	
OND 21	77	13.33	Neg	Headache 2/3
OND 22	46	4.17	Neg	

Abbreviations: Pos, positive; Neg, negative; NA, not available.
*samples with low sequence count automatically scored as -8.89

CHAPTER SIX

UNPUBLISHED RESULTS

AIM II: AGS is a unique feature of disease

Overview and rationale

Early work evaluating AGS score performance on single cell Sanger VH4 repertoires of CSF B cells in RRMS and OND patients identified a single subgroup of OND patients with consistently high AGS scores. These patients were diagnosed with NMO and thus excluded from all subsequent OND cohorts studies we performed to evaluate AGS performance using an NGS platform. However, this observation spurred an interest in expanding the use of the AGS tool to other diseases groups. Specifically, the AGS discovery method was first used against memory healthy control peripheral VH4 B cells, which focused the identification of RMF divergent codons on any that were skewed in RRMS. In this new application of AGS discovery, using two distinct disease-specific repertoires as training sets favors the identification of mutation patterns with targeted differential diagnostic potential.

The work presented below outlines the approach that was ultimately successful. It should be noted that many different modifications to the AGS design method outlined in previous research (Cameron et al., 2009) were tested as the original approach did not yield significant codons when applied to this fundamentally different control repertoire. The key difference in the new method is the “mutation unit” used. Previously, a single RM counted as one mutation event, whereas the updated method identifies “point mutation hotspots”, where multiple nucleotides mutated to cause a single RM each count

as mutation events. For example, a mutation of GGG to AGG is a RM that is weighted by one, whereas a mutation of GGG to AGT is a RM that is weighted by two. The various methods tested and implications of the success of the point mutation hotspot method will be further outlined in the Discussions section.

A UNIQUE ANTIBODY GENE SIGNATURE DIFFERENTIATES PATIENTS WITH NEUROMYELITIS OPTICA FROM THOSE WITH OTHER NEUROLOGICAL DISORDERS.

The following work has not yet been published due to limited numbers of NMO patients in all our published and unpublished NGS studies detailed in earlier chapters.

Introduction

Neuromyelitis optica (NMO) and multiple sclerosis (MS) are inflammatory demyelinating diseases that affect the central nervous system (CNS). NMO is typically characterized by an acute episode of inflammation and symptoms limited to the optic nerve and spinal cord, resulting in loss of vision and potentially leading to severed respiratory complications (Wingerchuk et al., 1999). In contrast, MS has a wider range of clinical manifestations and diagnosis relies on the relapse of demyelinating attacks or non-episodic disease progression over a longer period of time than NMO, as well as the detection of lesions in the brain (McDonald et al., 2001). However, these differences are not always apparent, as some patients with NMO will also present with brain lesions (de

Seze et al., 2003; Wingerchuk et al., 2006) and the brain lesions characteristic of MS vary among different patient subsets, such as in pediatric MS (Hahn et al., 2004). Rapid and early differential diagnosis of MS and NMO is essential since MS treatment such as interferon-beta therapy can actually increase the relapse-rate of NMO, for which delayed treatment can result in permanent damage to the CNS (Uzawa et al., 2010).

The role of B cells in the pathogenesis of both MS and NMO has been shown (Lucchinetti et al., 2002; Owens et al., 2006). B cells play a role in the pathogenesis of MS through antibody production (Lucchinetti et al., 2000; Antel and Bar-Or, 2006), antigen-presenting cell function (Harp et al., 2010; Ireland and Monson, 2011; Ireland et al., 2012) and cytokine production (Duddy et al., 2007; Ireland and Monson, 2011; Ireland et al., 2012; Ireland et al., 2014). Unlike MS, a unique NMO-specific autoantibody pool reactive to aquaporin-4 (AQP4) has been identified in many patients (Lennon et al., 2005). However, since AQP4 is detected in roughly 50-75% of patients with NMO or high-risk syndromes of the disorder (Lennon et al., 2004), identifying antibodies that are reactive to other antigens should improve the specificity of the diagnosis.

Our laboratory previously identified a pattern of somatic hypermutation in the VH4 antibody genes of CSF-derived B cells that are not found in other neurological diseases (OND). The presence of this antigen gene signature (AGS) in the cerebrospinal fluid (CSF) of clinically isolated syndrome (CIS) patients was demonstrated to predict conversion to definite MS with 91% accuracy (Cameron et al., 2009), and subsequently confirmed to be present in MS brain tissue (Ligocki et al., 2010). Because B cells in the CNS contribute to NMO pathology, we hypothesized that B cells in the CNS would carry

antibody genes with somatic hypermutation patterns not found in healthy donors or OND patients. Our analysis isolated a distinct gene mutation pattern in VH4 NMO CSF B cells that distinguishes them from MS and OND B cells.

Methods

B cell repertoire analysis

Antibody repertoires at UT Southwestern Medical Center (UTSWMC) were generated from singly sorted CD19+ CSF B cells using single cell PCR as previously described (Monson et al., 2005b; Harp et al., 2007) and in accordance with the UTSWMC Institutional Review Board (IRB). Antibody repertoires from University of Colorado Denver (UCD) were generated from singly sorted CD19+ and CD138+ CSF B cells (Bennett et al., 2008). Single cell-sorted CD19+ peripheral blood (PB) B cells were also included in the analysis (Brezinschek et al., 1995). Antibody repertoires from patients were separated into three groups based on clinically defined conversion to MS (Geurts et al., 2005) or NMO (Wingerchuk et al., 2006). The third group consisted of patients with OND and healthy controls (HC).

Mutation analysis

VH4 sequences with one or more replacement mutations (RM) were analyzed. All sequences were aligned using IMGT/V-QUEST (Lefranc, 2003). This information was subsequently analyzed using a Perl-based program and the codons numbers were converted to the Chothia numbering system (Chothia and Lesk, 1987; Al-Lazikani et al., 1997). Any sequence that was non-productive, misaligned or incomplete between codons

31 and 103, to include all the complementarity determining regions (CDR), was discarded. The analysis was restricted to codons 31-92 to avoid including the hypervariable CDR3 region.

Signature Identification

VH4 gene signature was determined by identifying point mutation hotspots in the NMO patient-derived sequences compared to those from the MS cohort (filtered as described above). To ensure the biological relevance of the hotspots identified, only point mutations found at codons with a RM (hereafter referred to as events) were counted. A hotspot was defined as a codon with a significantly elevated (p-value < 0.05) event frequency in the NMO sequences, as determined by Pearson's chi-square test using the MS sequences to establish expected event frequencies. This analysis revealed 12 VH4 codons (36, 39, 45, 46, 50, 59, 61, 65, 67, 70, 86, 90) that were significantly targeted in NMO. Figure 6-1 provides a side by side comparison of a VH4 protein structure and AGS codon positions between MS AGS (6 codon) and this new NMO AGS (12 codons).

Patient Scoring

The identified AGS in NMO sequences (NMO-AGS) was scored for each patient based on events per 100 sequences. Only patients with 6 or more sequences were evaluated. The threshold NMO-AGS score to distinguish the NMO and MS patient cohorts was defined as the average + 2 S.D. NMO-AGS of the MS patients (equal to or greater than 120).

Results

The VH4-specific NMO-AGS enabled the separation of the NMO patients from the MS and OND groups (Figure 6-2). The AGS identification method restricts the inclusion of codons in the mutation signature to those that do not excessively fluctuate in RMF in the control group (MS), as the mean + two standard deviations of the MS scores did not overlap with the lowest scoring NMO patient. The non-NMO OND patients also had low NMO-AGS scores, despite not being included in the training repertoire, reinforcing the observed differences in mutation profiles between NMO repertoires and other OND repertoires previously observed using the MS AGS tool.

Since the NMO-AGS score is dependent on RM count at specific codon positions per 100 sequences, as well as the number of point mutations that cause these RMs, we evaluated any differences in RMF between cohorts to ensure observed score differences were not a result of diverse RM counts per sequence. The NMO repertoire had an average RMF per sample only 11% greater than that of the MS repertoire (Table 6-1), in contrast to a greater than two-fold increase in NMO-AGS scores compared to MS. We also noted that this small difference in RM per sequence was maintained when individual point mutation counts that contribute to the RMs were factored in and used for scoring. In fact, all 3 cohorts had a sample average of 1.2 point mutation events per RM, thus ensuring that the only significant impact on score was RM position.

Testing of Illumina sequencing for future implementation

Overview and rationale

As outlined in the introduction, the sequence length of Illumina technology data output has been steadily increasing as improvements to the platform have yielded a decrease in sequence quality drop-off in function of sequence length. The marked transition from a maximum of 150 base pair reads (Loman et al., 2012) to the currently available 250 base pair reads makes this platform potentially viable for full-length VDJ segment sequencing required for MSPrecise scoring and genetic analysis of BCRs. As the average length of the sequences of interest is roughly 350 base pairs long (with variations due to a range of possible CDR3 lengths from several to over 100 nucleotides), by using paired-end reads with Illumina sequencing, full sequence coverage can be obtained.

Results

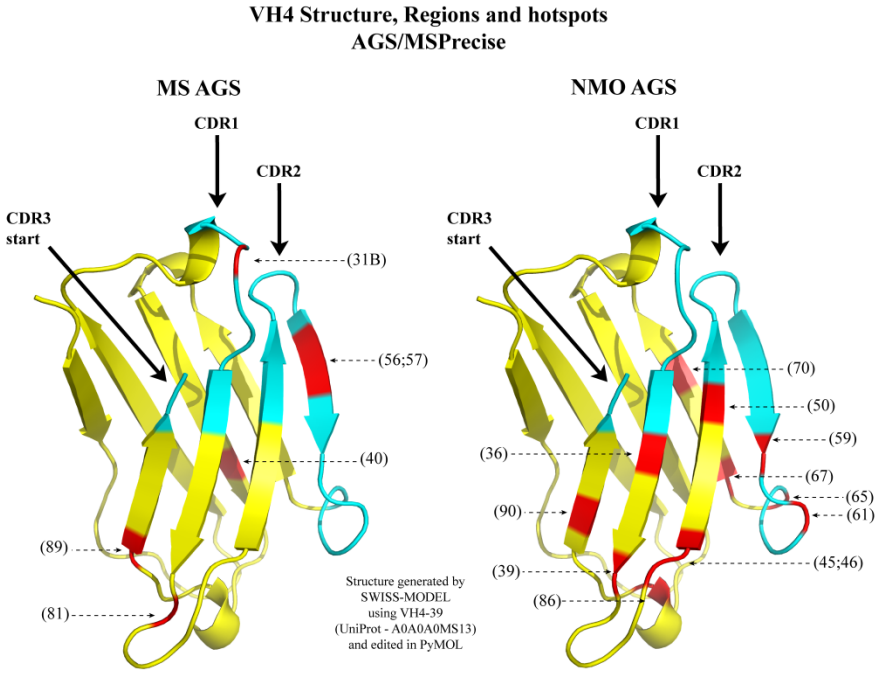
Three healthy peripheral naïve 1000 B cell sequence libraries were prepared with the BIOMED2 FR1 PCR primer set and sequenced (Figure 6-3). One of the challenges encountered with the BIOMED2 PCR protocol, was obtaining successful PCR amplification using pooled primers. After multiple rounds of testing, it was determined that amplification was more successful with 4 separate reactions: VH1 alone, VH3 alone, VH4 alone and VH2, VH5, VH6/7 (one primer) combined. Successful reactions that were still detectable by gel after adapter addition would then be combined after gel purification prior to sequencing. VH2/5/6/7 primers were combined to increase the likelihood of a successful PCR reaction for these low abundance families. Nevertheless, the VH2/5/6/7

reaction was the only one that failed frequently. As shown by the VH family distribution data (Figures 6-3, 6-4), the VH2/5/6/7 reactions did not end up yielding productive sequence reads as determined by sequencing output alignment.

The other key evaluation of the Illumina platform was to identify the % of sequence coverage attributed to crossover (CO) reads, i.e. reads with shared CDR3s across patients (Figure 6-5). The data matched observation from 454 sequencing that attribute CO reads to highly amplified sequences that have outliers associated with other samples than their source. This is indicated by the very high sequence coverage of a small number of clones with over 99% or reads belonging to one sample (difference between the green and red lines in Figure 6-5). When CO removed reads are not removed from the sample they are associated with at over 99%, CO sequence coverage is found to be extremely low by Illumina sequencing.

FIGURES FOR CHAPTER SIX: UNPUBLISHED RESULTS

Figure 6-1. VH4 protein structure and hotspots. VH4-39 protein structure obtained through SWISS-MODEL (Arnold et al., 2006; Bordoli et al., 2008; Biasini et al., 2014) is shown in duplicate, with either the MS AGS codon hotspots highlighted in red (left; 6 codons) or the NMO AGS codon hotspots highlighted in red (right; 12 codons). All hotspots are marked with Chothia numbers. CDR regions are highlighted in cyan. Only the VH4 germline portion of CDR3 is included in the protein structure, which is 2 amino acids long. Protein structure was edited in PyMOL (Schrodinger, 2015).



<u>Figure 6-2.</u> The NMO AGS clearly separates the NMO and MS training cohort, and also distinguishes between NMO and non-NMO OND patients. Each point represents one patient; mean and standard deviation (SD) are also indicated. The cut-off point of 120 (MS mean + 2 SD) is indicated by a dotted line. Statistical significance of the difference between cohorts was calculated by Mann Whitney test.

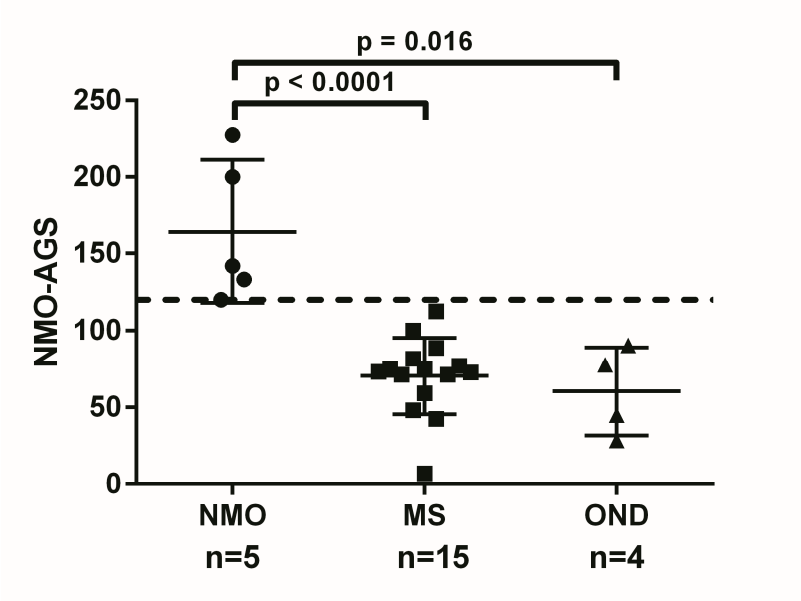
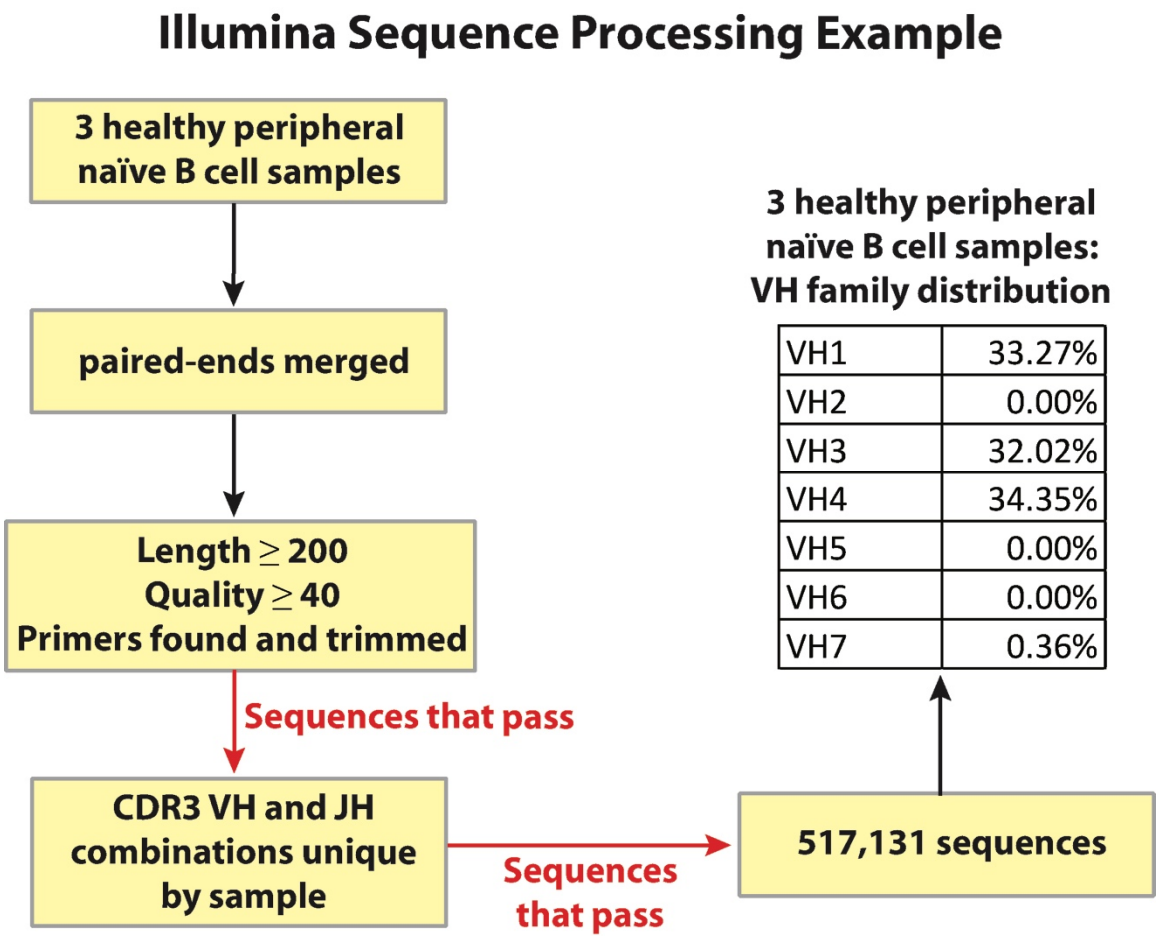
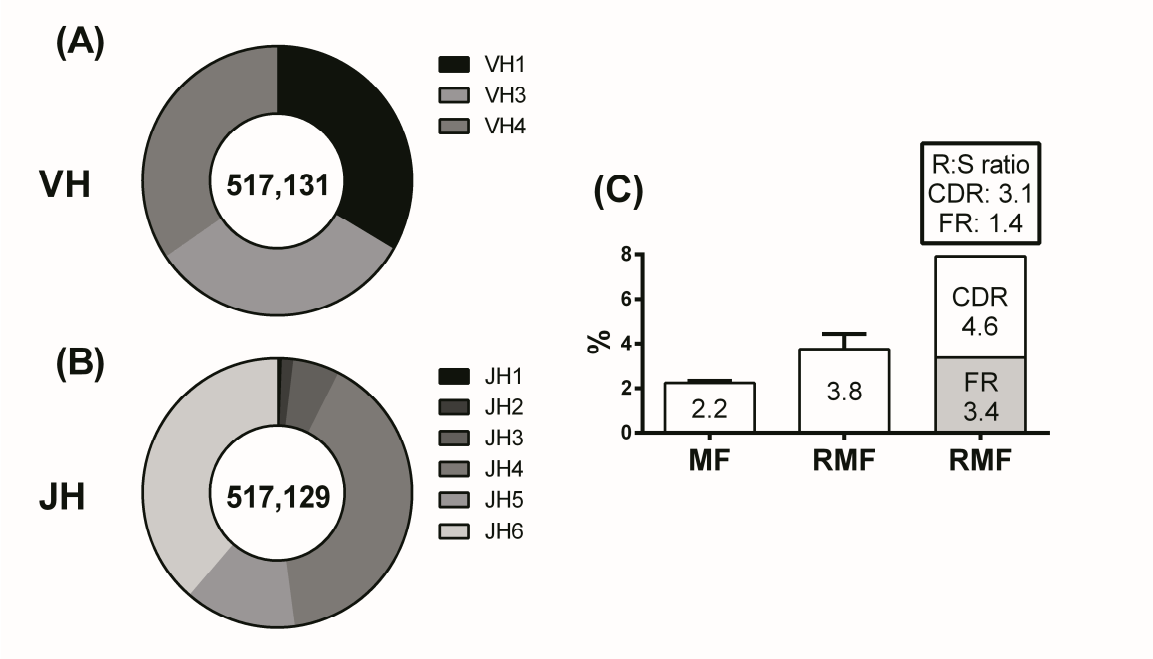


Figure 6-3. Illumina sequencing analysis summary. All sequence counts and percentages are based on unique sequence counts, as defined by CDR3 amino acid, VH gene and JH gene match.



<u>Figure 6-4.</u> Illumina sequencing analysis. All data is based on unique sequence counts for 3 naïve healthy control peripheral blood B cells.



<u>Figure 6-5.</u> Illumina unique sequence clone analysis. Clones (unique CDR3 amino acids) were sorted in order of read coverage and are shown here as % sequences that belong to a clone that is found in more than one patient (green) or with the standard correction to remove the crossover (CO) label from reads that represent more than 99% of a clone and belong to a single patient (red).

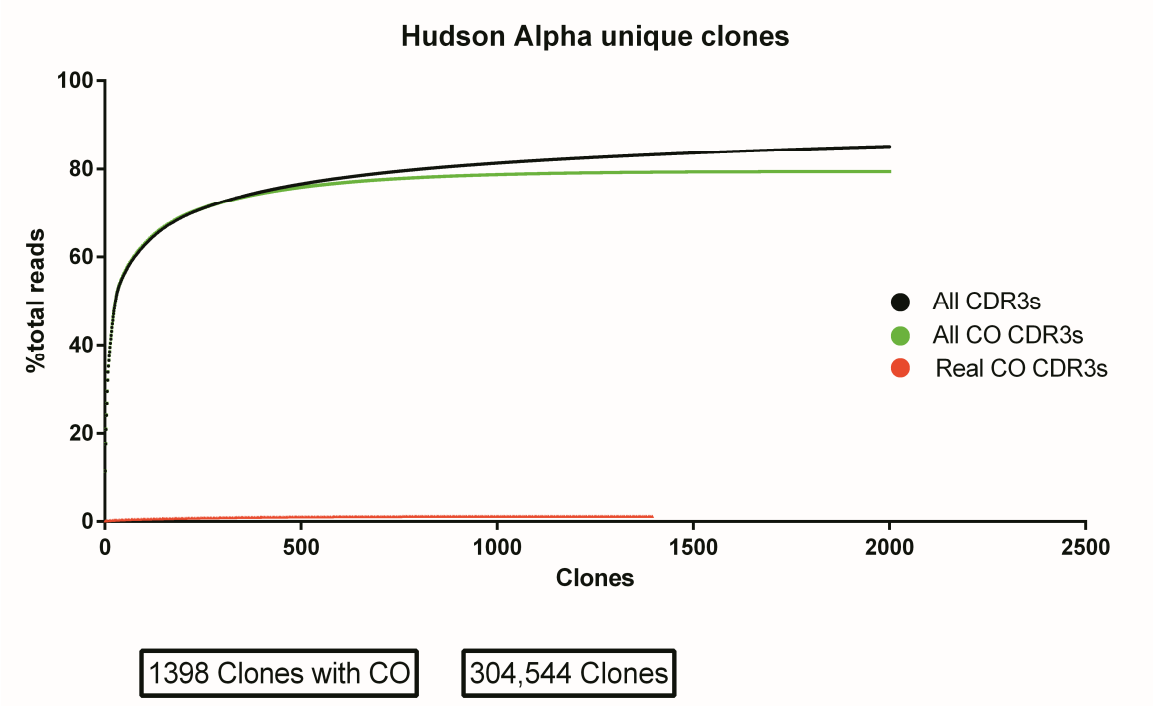


TABLE FOR CHAPTER SIX: UNPUBLISHED RESULTS

Table 6-1. VH4 sequence count and RMF in NMO, MS and OND patients. The number of VH4 sequences with at least one RM (used for NMO-AGS score calculation) is indicated for each patient. RM frequency (RMF) indicates the % of codons 31-92 that have RM in each patient.

		VH4 seq w/RM	VH4 RMF
MS	**CIS831**	58	11.3%
	ON3-05	49	12.7%
	M522	44	11.3%
	CIS348	41	12.4%
	M357.1	40	8.2%
	M584	22	13.0%
	M125	21	11.6%
	ON5-02	18	12.9%
	MS3-01	16	14.6%
	CIS431	14	11.5%
	ON4-07	12	10.8%
	ON3-03	11	13.0%
	ON4-08	11	10.0%
	M368	8	13.9%
	CIS429	7	13.7%
	Average	**24.8**	**12.1%**
NMO	**ON07-5**	31	10.7%
	ON09-9	15	9.2%
	ON10-1	15	18.2%
	TUM-527	11	14.0%
	ON09-3	10	16.6%
	Average	**16.4**	**13.7%**
OND	**BF2N**	22	9.4%
	ON3-01	21	8.5%
	CIS563	20	18.2%
	M341	9	11.8%
	Average	**18.0**	**12.0%**

CHAPTER SEVEN

DISCUSSION

NGS data features and challenges

The most challenging aspect of NGS usage is the impact of errors on the observed sequence data output. With Sanger sequencing data, which is generated by averaging sequencing signal of all the templates in a single reaction, if at least 80% of the templates in a reaction are identical, the final single sequence output is correct (Davidson et al., 2012). In contrast, for NGS, all templates with a sequencing primer attached generate a sequence read. As a result, there are many different types of errors that will affect the final output.

These errors come in two types for a standard NGS protocol: process errors and sequencing errors. Process errors refer to those that occur as a direct or indirect result of the sequencing library preparation protocol. Specifically, issues such as amplification bias, DNA polymerase error rates and sequence contamination prior to barcode attachment all have a direct impact of the sequencing repertoire after NGS. Instead, sequencing errors are sequencing platform and technology dependent, and refer to errors generated as a result of signal detection or interpretation errors by the sequencing platform and/or corresponding software.

Historically, NGS use has focused on VDJ recombination gene selection to identify clonally related B cells (Boyd et al., 2009; Arnaout et al., 2011; Logan et al., 2011). In this context, the impact of background nucleotide mutations is minimized, since alignment to a germline repertoire is the primary goal. In fact, we were one of the first groups to use NGS to study somatic hypermutation (SHM) in real patient samples in our

work prior to and published in 2014 (Rounds et al., 2014). Because of this, many of the challenges we faced in transitioning from Sanger to NGS sequencing were completely new at the time. For instance, the effects of barcode contamination on NGS repertoire analysis accuracy, which will be discussed further below, had never been openly published or discussed prior to our encounter of it in the context of the Verification study. Fortuitously, our initial forays into NGS data analysis included detailed cross-checking across samples of identified B cell clones as legacy of single-cell PCR quality testing methods which test for high risk of DNA contamination in these low starting template reactions. These initial observations introduced us to the importance of unique sequence and clone definitions in NGS repertoires, which have only recently been comprehensively evaluated by others in the field (Jiang et al., 2015; Yaari et al., 2015; Khan et al., 2016). To this effect, the combined use of unique molecular identifiers, which tag single DNA templates, with spike in standard BCR templates has shown that PCR amplification and sequencing can yield almost 100-fold more "unique" CDR3 AA sequences from a starting control pool of 16 (Khan et al., 2016). As a result, the application of NGS to unique sequence and SHM determination in the studies presented here has focused on identifying better methods to reduce and compensate for both process and sequencing errors, which will be discussed here.

Error rates and sources

The studies presented in Chapters III through V were the subject of incremental sequencing pipeline changes between each study as a result of sequence repertoire evaluation and changes in available technology and reagents. For example, while

Chapters III and IV used Phusion DNA polymerase (NEB), Chapter V introduced Q5 DNA polymerase as the choice for PCR, a new addition to NEB's collection of DNA polymerases with a reported lower error rate of 2-6 fold that of Phusion. While such improvements to DNA polymerase fidelity are dependent on factors outside laboratory control, the way the polymerase is used in sequencing library preparation is open to significant optimization with regards to amplification cycle number. In fact, the impact of error rate is directly proportional to template size and amplification cycle number. The other effect of PCR amplification, sequence coverage bias, will be discussed in the next section.

In order to illustrate the impact of PCR reagents and cycle number on error rates, here is a comparison of the Verification study (Chapter IV) and Validation study (Chapter V) methods:

- Verification = Phusion DNA polymerase with 23 cycles external PCR + 31 cycles nested PCR + 10 cycles barcoding PCR (Rounds et al., 2015)
- Validation = Q5 DNA polymerase with 30 cycles target PCR + 10 cycles barcoding PCR

Template DNA in the Verification study thus went through 1.6 fold more amplification cycles than its counterpart in the Validation study. Using the lowest reported fidelity of Q5 compared to Phusion (2X higher) and the NEB-reported Phusion DNA polymerase error rate of $4.4*10^{-7}$ errors per base pair, the Verification study would have an error rate of $2.82*10^{-5}$ errors per base pair and the Validation study one of $8.8*10^{-6}$ with the strongly error-minimizing assumption that each error is only amplified once. If only the impact of nucleotide mutation in the somatic hypermutation region of interest is

considered, i.e. codons 31-92 (with a maximum of 67 codon positions, so roughly 200 base pairs in length), that represents $5.64*10^{-3}$ errors per sequence in the Verification study (1 in 177) and $1.76*10^{-3}$ errors per sequence in the Validation study (1 in 568). To put that into context, the average number of raw sequences (filtered only for length and quality by VDJpipe) in the samples from the Validation study was 2300. Again, these numbers ignore the amplification of templates with errors that occurs after the mismatch event and provide a range for the expected number of distinct PCR errors in function of total sequences observed.

With those numbers in mind, the sequencing error rate of the 454 Titanium platform (Roche/454, Branford, CT) can also be factored in. The error rate of this platform has been extensively studied and found to predominantly homopolymer length overestimation or underestimation based (Balzer et al., 2010; Bolotin et al., 2012; Loman et al., 2012; Georgiou et al., 2014). In fact, because the sequencing cycles occur for one nucleotide at a time, the substitution error rate is essentially considered to be 0 since any observed substitutions on non PCR amplified DNA are a result of homopolymer overestimation followed by a different homopolymer underestimation, or vice versa (Balzer et al., 2010). The reported homopolymer error rate for the 454 Titanium platform is $3.8\text{-}5*10^{-}3$ errors per base pair (Loman et al., 2012; Georgiou et al., 2014). This translates to an average of one insertion or deletion sequencing error per sequence. In practice, due to stochastic variation, about one in three sequences has observed insertions or deletions (31% or reads in the Validation study). Contrary to DNA polymerase-induced errors, insertion deletion sequencing errors are easily identifiable by alignment

data analysis. The primary consequence of removing sequencing errors from a 454 NGS repertoire is the significant reduction in sequence coverage that this results in.

Amplification and the unique sequence

In sequencing repertoire analysis, the identification of a "unique" sequence read is critical because it impacts all antibody genetic analysis metrics. The most common approach when working with PCR-amplified data is to collapse all sequences with exact nucleotide matches down to a single sequence with an associated sequence coverage value. The benefit of this approach is to limit the impact of amplification bias on repertoire analysis of features such as gene alignment frequencies. In the case of the Validation study for example, sequences ranged from having no exact nucleotide matches ranged (a sequence coverage of 1) to over 14,000 matches.

However, there are several limitations to solely condensing unique reads by exact nucleotide matches as was done for the Confirmation study (Chapter III). With this method, any incidence of PCR or sequencing error rates is guaranteed to produce one additional "unique" read per error (Figure 7-1). Additionally, complementary insertion and deletion events that occur in close proximity only frameshift the sequence in the area between these events, thus making them difficult to identify if they occur in the parts of the sequence that do not have germline region to align to, i.e. the CDR3 region (Figure 7-1). Another source of "unique" reads that are added due to errors was observed in work for the Verification study (Chapter IV), where sequences with primers trimmed were found that matched each other with the exception of several nucleotides at either edge of the sequence. This was problematic, since only exact nucleotide matching sequences

were condensed into a single output sequence, so an allowance of up to 5 nucleotide mismatches total from either primer trimmed sequence edge was allowed for 2 sequences to be considered as matching.

For the Validation study (Chapter V), several changes were implemented to address the issue of PCR-generated errors which artificially expand the pool of unique sequences that are identified by repertoire data analysis. In addition to the significant changes to PCR cycle counts and DNA polymerase updates previously discussed, a new definition of a unique sequence was implemented. Unique reads were defined by VH and JH gene alignment, as well as their CDR3 nucleotide sequence and RM pattern, rather than by exact nucleotide matching across the whole length of the read (with the exception of edge mismatches). By requiring only RM matches as defined by their position, germline nucleotides and sequence nucleotides, PCR errors that generated silent mutations would be ignored and not result in more "unique" sequence counts. The other effect of this new unique sequence definition was to avoid the issue of sequence edge mismatches that had to be specifically addressed previously (Chapter IV), since these edge mismatches occur outside the 31-92 numbered codon range where RMs are counted.

The impact of unique sequence definition on repertoire diversity has been evaluated for each study. One way to measure it was to look at the diversity index (DI) which quantifies relative abundance of VH4 genes. The median DI for RRMS patients changed from 0.6 in the Verification study to 1.2 in the Validation study, where higher scores mean more diverse representation of VH4 genes. This suggests that the process changes and/or unique sequence definition changes implemented in the Validation study compared to the previous work reduced repertoire skewing towards specific gene

alignments. Coverage ratio was another metric used to help capture amplification biases in the repertoires and first introduced in the Validation study. Instead of gene alignment diversity, it focuses on high sequence amplification outliers by representing the ratio of sequence coverage for the 2nd most abundant sequence to the sequence coverage of the most abundant sequence in a single sample. For example, in a sample with equal sequence coverage for all unique reads, the coverage ratio would be 100%, whereas a coverage ratio of 1% indicates that 1 unique sequence matches 100 more raw sequence reads than the next most amplified unique sequence. In the Validation study, coverage ratio and DI were found to be connected (Figure 7-2), with a very large number of samples below 2% sequence coverage having low DI (more than 75% below 1.0) compared to samples with even only 2-5% sequence coverage (50% over 1.0; $p = 0.018$) with this gap growing as coverage ratios increase to 100%. As a result, samples with less than 2% coverage ratio were required to have 3 times more unique reads than those with higher coverage ratios in order to pass the low sequence sample filter.

Sequence "crossover", "crosstalk" or "barcode contamination"

Another challenge of NGS data generation is the incidence of "crossover sequences", a term we use to describe sequences with CDR3s that match between samples from different patients. The advent of NGS has allowed for unprecedented deep sequencing of the human BCR repertoire in peripheral blood. This has led to the confirmation that finding matching CDR3 subsequence between two different patient repertoires is an extremely rare event (0.13%) (Arnaout et al., 2011). In the context of BCR sequencing from CSF B cells, expected numbers of unique templates in the

hundreds suggest that any occurrence of matching CDR3s in these studies are likely due to this contamination issue. Over the course of the studies, it was observed that this contamination was often tied to highly amplified templates from a single sample being found with low abundance in another. This has led to the 99% CO filtering rule that is used to identify the sample of origin for a specific CO sequence and only remove it from the other samples it contaminates (Figure 7-3), to avoid filtering out sequences with the most coverage of the entire repertoire (see Figure 6-4).

Surprisingly, publications acknowledging this issue are still quite rare (Quail et al., 2014; Seitz et al., 2015) and is only now being openly addressed by biotech companies ("some of the data produced by this method exhibited alarming levels of cross-contamination (crosstalk) between the barcodes" IDT in the context of the TruGrade™ Processing Service).

Alignment error

The greatest single requirement to proper evaluation of both clonal diversity and SHM is the accuracy of the reported gene alignment. Without this, any optimizations to process and sequencing errors as well as unique sequence definitions are limited in their ability to correct perceived errors in the final NGS repertoire. In this context, the accuracy of the most widely used BCR germline database, IMGT, is critical. Although NGS has allowed for evaluation of the completeness of the IMGT germline reference set with regards to new polymorphisms (Wang et al., 2011; Watson and Breden, 2012; Gadala-Maria et al., 2015), a corresponding thorough investigation into the validity of the original references has yielded strong evidence that 100 of the 226 IGHV alleles are almost certain to include errors and should be removed (Wang et al., 2008; Boyd et al.,

2010). To summarize these findings, alleles were found without published reference data or with only one source, with truncated ends (i.e. not a full germline sequence), from a source that found more than 2 alleles per individual, not aligned to a comprehensive rearranged VDJ sequence database from EMBL and/or not aligned to any sequence from newly generated NGS data on BCRs from peripheral blood. The impact of germline reference data set on MSPrecise® alignment cannot be overemphasized, since mutations are dependent on sequence comparison to germline.

MSPrecise®

Previous work on the association of MSPrecise® with disease has shown that the AGS is found in the CSF of RRMS patients and CIS patients who will convert to CDMS (Cameron et al., 2009) as well as in the CNS of MS patients (Ligocki et al., 2010). Clones that are found in both the CSF and CNS have also been identified in patients with MS (Obermeier et al., 2011). Furthermore, the finding that AGS-enriched antibodies from both CIS and RRMS patients exhibit binding to neurons and astrocytes in the brain (Ligocki et al., 2015) also suggests that AGS-positive B cells are associated with disease.

Most of these studies associate elevated MSPrecise® scores with patients at the early stages of disease. As a result, a better understanding of MSPrecise® changes over the course of disease progression is highly valuable in order to start addressing the question of AGS-positive B cells' role in MS. One of the challenges of assessing MSPrecise® in patients with long-term RRMS is the potential effects of disease modifying therapies (DMTs) on score. In the Verification study (Chapter IV), 5 patients on DMTs for at least 9 months and with CDMS for over 2 years were included.

Surprisingly, despite an 84% accuracy for MSPrecise® in this study, all long-term patients with sufficient repertoire sequence counts had MSPrecise® scores over the CDMS threshold (Rounds et al., 2014). The one patient with low sequence count in the repertoire, a feature we found to be associated with OND and use as benchmark for assigning low score to patients, was on Natalizumab for 4.5 years, which blocks B cell entry in the CNS. Although we currently lack significant evidence to connect specific DMTs with changes in MSPrecise® score over time (or lack thereof), this data suggests that NGS sequence repertoire size is associated with B cell counts in the CSF, despite the challenges of repertoire diversity overestimation as a result of process and sequence errors.

Illumina versus 454 sequencing

As mentioned in Chapter VI, there are several notable differences between 454 and Illumina NGS technology. While Illumina MiSeq can provide roughly 10-fold greater sequence coverage compared to current 454 sequencing, it also comes with an increase in substitution error rate up to 1.5%, or on average about 3 per BCR sequence in our repertoire (Fuellgrabe et al., 2015). As a result, the use of Illumina for SHM evaluation has not currently been tested, although existing research on error correction from bacterial genome analysis studies have explored the use of paired-end assemblers for Illumina error correction (Schirmer et al., 2015). Another important feature of error rates for Illumina sequencing is that they closely correlate with read quality and sequence length as they increase with the accumulation of DNA molecule elongation failure events (Schirmer et al., 2015). Currently, VDJServer requires a minimum alignment of 10

nucleotides on a local alignment score in order to reassemble paired-end sequence reads. If a match is found, forward and reverse sequence nucleotides downstream or upstream (respectively) of that alignment are removed to generate a full-length sequence output. As a result, this acts as a form of error correction by effectively trimming over 50 nucleotides of the lowest quality from each paired forward and reverse sequence before assembly (for 250 base pair long paired-end reads and an expected full sequence lengths of around 350).

FIGURE LEGENDS FOR CHAPTER SEVEN: DISCUSSIONS

Figure 7-1. PCR and sequencing errors' impact on sequencing results. This figure represents the possible sequencing results obtained from a single starting template sequence. DNA polymerase errors are marked in red and sequencing errors are marked in purple. Result sequences labelled with * indicate those that are easy to remove based on detectable frameshifting due to single insertion or deletion events. Sequences with an equal number of insertion and deletion events are only easily identified if these events occur in a germline aligned portion of the sequence (in the CDR3 for instance, this would not be clear).

Figure 7-2. Diversity index distribution by coverage ratio cut-offs. All RRMS and OND samples that pass quality filters from the Validation study (including samples marked as having low sequences; N=97) are each represented by a single DI value, grouped by coverage ratio ranges. Median and interquartile ranges are marked on the figure. Samples with less than 2% coverage ratio were required to have 3 times more unique reads than those with higher coverage ratios in order to pass the low sequence sample filter.

Figure 7-3. Crossover sequence removal. CDR3 sequences are color-coded by sample of origin. CDR3 sequences are shown at each step of the pipeline: starting gDNA template (1 copy), PCR amplified and sequenced (grouped by sample = black circle),

sequence filtered to remove CO CDR3s, and collapsed down to unique sequences prior to genetics analysis and MSPrecise® scoring.

FIGURES FOR CHAPTER SEVEN: DISCUSSION

Figure 7-1.

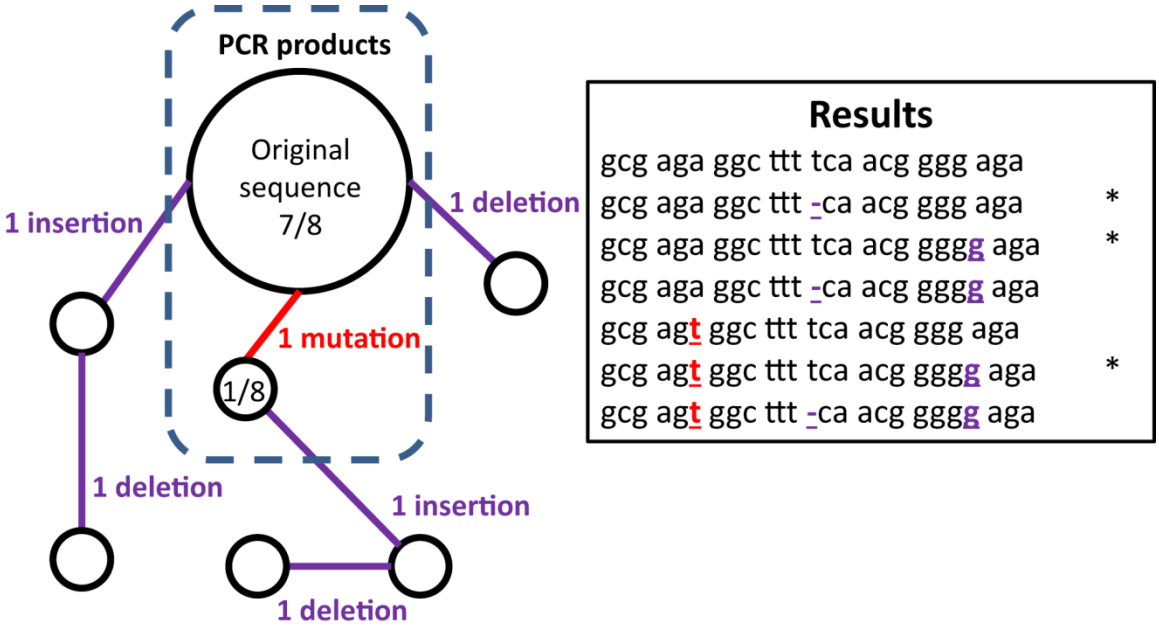


Figure 7-2.

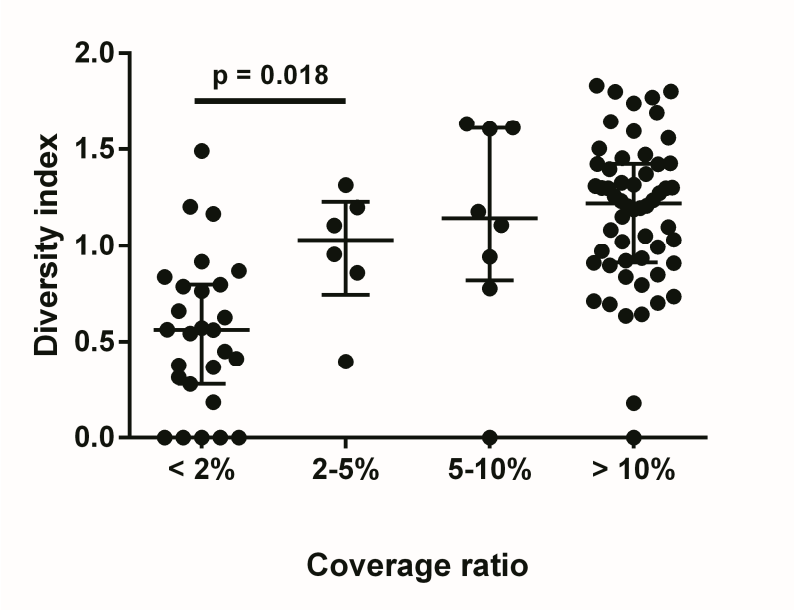
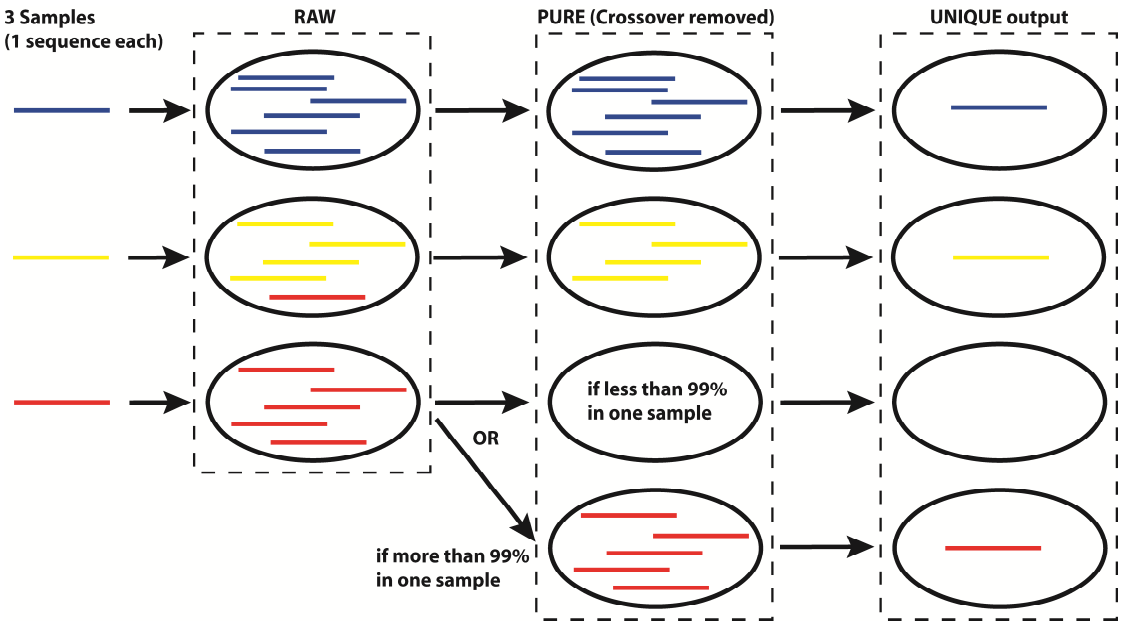

p = 0.018
Diversity index
2.0
1.5
1.0
0.5
0.0
< 2%
2-5%
5-10%
> 10%
Coverage ratio

Figure 7-3.


3 Samples
(1 sequence each)
RAW
PURE (Crossover removed)
UNIQUE output
if less than 99%
in one sample
OR
if more than 99%
in one sample

CHAPTER EIGHT

FUTURE DIRECTIONS AND CAVEATS

Repertoire analysis

The transition from single-cell sorting-dependent BCR analysis to CSF cell pellet derived NGS repertoire generation has enabled a more detailed and precise characterization of MSPrecise® and its clinical diagnostic potential. In breaking new ground by evaluating the feasibility of SHM pattern evaluation by NGS, we encountered many challenges that have not been addressed by scientists in the field of immune repertoire analysis until recently. Incremental changes to sequence library preparation and sequencing were implemented as a result of performing multiple distinct and sequential studies on CSF BCR sequences. Although this allowed for more optimization between studies, which was necessary due to a lack of standard practices for these new techniques, it limits our ability to compare results across the 3 studies (Chapters III-V). Nevertheless, the combined studies highlight the single most impactful library preparation feature for repertoire analysis: amplification bias. A recent study using synthetic antibody genes combined with unique molecular identifiers demonstrated the severe impact that amplification can have on estimates of repertoire diversity from a small number of unique starting templates (Khan et al., 2016). In the Validation study, the focus on unique sequence definition and amplification bias (as measured by the coverage ratio) highlights a minimum coverage ratio for which diversity indexes begin to fluctuate wildly, suggesting that low unique template number is related to greater potential for over amplification.

One of the main issues with the sequence data curation methods detailed in these MSPrecise® studies is that they are dependent on observations made at the cohort level. In reality, PCR amplification varies from sample to sample, so some groups have tested various methods of including control sequences or sequence elements into each sample PCR reaction to track and be able to correct process and sequence errors during repertoire analysis. One approach is to use "molecular amplification fingerprinting", a two-step unique molecular identifier incorporation strategy with one identifier incorporated before PCR amplification in a single reverse-transcription step, and one added at the very end (with sample barcode and sequencing primers) (Khan et al., 2016). This allows for an evaluation of amplification bias based on abundance of unique identifier pairs, and for and evaluation of errors within the cluster of sequences that belong to a specific starting template, thus reducing overestimation of sequence diversity in a sample. The other advantage of this approach is to minimize the over correction of sequence diversity that occurs with any of the unique sequence definition methods implemented in the studies, which results in a single sequence output being generated for one or more identical BCRs in the original cell pellet. Another use of the unique sequence inclusion approach is the use of synthetic antibody gene sequences that are used for error rate testing (Quail et al., 2014). In this case, these uniquely barcoded sequences serve as a monitoring tool which can identify not only incidence of barcode contamination, but also sequence error rates. The main limitation for this approach is that it emphasizes error detection over correction, and barcode contamination (i.e. crossover) detection between patients is relatively straightforward with low cell count BCR sequencing from the CSF.

Overall, the main limitation of these corrections methods is the inability to offset the impact of RepliG amplification of the gDNA templates prior to PCR amplification. As a result, another area of sequence data correction that can be improved is the use of new tools for large scale repertoire analysis of clonal populations and lineage trees, such as Change-O (Gupta et al., 2015) which is currently being incorporated into the VDJServer tool suite. Another useful feature of this tool is the determination of the Ig variable region gene segment alleles carried by a specific patient, which can help correct for the excess alignment diversity caused by background errors and which can result in alignments to more than 2 alleles of a specific gene per patient sample.

MSPrecise®

As detailed earlier, amplification bias directly impacts MSPrecise® score since over amplification of sequences with AGS positive or negative RMs will change the score if not corrected. Over amplified sequences are more likely to accumulate process and sequencing errors, and thus have a greater impact on score. This is why we cannot rely on raw sequence data alone and why the updated definition of a unique read is so critical. Improvements to unique sequence and clone determinations will thus bring MSPrecise® scores closer to their precise value, i.e. the exact RM pattern score if each productive BCR in the CSF cell pellet is counted once.

Other areas of improvement for MSPrecise® characterization are the determination of score fluctuation in function of disease course. While early onset MS patients at the point of relapse have been extensively studied and scored, fluctuations in score during phases of remission are still unknown. This is largely due to the challenges

of obtaining CSF samples from patients who are not experiencing or have not recently experienced an MS attack. One potential way around this issue would be to study whether MSPrecise® scores are reflected in peripheral B cells, which would greatly lower the burden of MSPrecise® testing for the patient. It is likely that B cells expressing AGS-enriched BCRs in the periphery will be lower in abundance compared to non-enriched B cells than their CSF counterparts. In this context, the use of techniques to correct amplification bias, such as with unique molecular identifiers, will be crucial to properly evaluate AGS positive B cell frequency in peripheral blood.

In addition to testing MSPrecise® fluctuation over relapse and remission phases, long-term evaluation of MSPrecise® scores over the course of treatment and upon conversion to progressive MS could yield valuable insights into the biology behind this mutation pattern and the potential role that AGS positive B cells are playing. One of the most valuable features that researchers and clinicians look for in a disease biomarker is correlations between biomarker levels and patient response to treatment. In the case of RRMS, marked by phases of subclinical CNS damage during which the patient does not know whether their treatment is working or not, such a connection between MSPrecise® score and response to a DMT (as measured by EDSS, new lesion formation by MRI, relapse rate and/or conversion time for CIS patients) would be invaluable. The evaluation of MSPrecise® scores compared to clinical markers of damage could also shed light into whether AGS positive B cells participate in autoreactivity through pro-inflammatory cytokine secretion or autoimmune antibody production.

APPENDIX 1

Perl script to convert VDJServer output to SQLite database

File generation: raw data files were assigned a unique two digit number inside each batch. PERL script descriptions are preceded with "#".

Folder organization: Parent folder (run here) → TSVs → 1-6 (each batch gets a numbered folder) → store the rc_out.tsv files (repertoire characterization output from VDJServer) and the dup.tsv files (VDJpipe output from VDJServer) for a single batch here.

```
# Last update = 04/29/16
# Code Writer: William H. Rounds
# Copyright UT Southwestern

use strict;
use warnings;
use Math::Trig;
use DBI;

# Start 6 way fork (one parallel fork per batch)
my @childs;
for ( my $f = 1; $f < 7; $f++) {
	my $pid = fork();
	# parent
	if ($pid) {
		push(@childs, $pid);
		print "Started process $f!\n";
	}
	# child
	elsif ($pid == 0) {
		#create error output
		my $outputfile = "Batch" . $f . "_SampleErrors.txt";
		open(ERR, ">$outputfile") or die("Error: cannot open file $outputfile");

		my $database = "Validation" . $f . ".db";
		unlink $database; #delete to avoid overwrite errors
```

```
my $dsn = "DBI:SQLite:dbname=$database"; #create new database
my $dbh = DBI->connect($dsn, { RaiseError => 1 }) or die $DBI::errstr;

#create SQL table for sequence data
$dbh -> do( "CREATE TABLE Sequence (
batch INT NOT NULL,
sample INT NOT NULL,
ID TEXT NOT NULL,
outofframe TEXT NOT NULL,
missCYS TEXT NOT NULL,
missTRP TEXT NOT NULL,
stopcodon TEXT NOT NULL,
indels TEXT NOT NULL,
framepreservingindels TEXT NOT NULL,
vgene TEXT NOT NULL,
jgene TEXT NOT NULL,
dgene TEXT NOT NULL,
cdr3 TEXT NOT NULL,
cdr3nucl TEXT NOT NULL,
agsRM TEXT,
agsSM TEXT,
agsRMnucl TEXT,
agsSMnucl TEXT,
CDR1l INT,
CDR1mut INT,
CDR1RM INT,
CDR1SM INT,
FR2l INT,
FR2mut INT,
FR2RM INT,
FR2SM INT,
CDR2l INT,
CDR2mut INT,
CDR2RM INT,
CDR2SM INT,
FR3l INT,
FR3mut INT,
FR3RM INT,
FR3SM INT,
cFR3mut INT,
cFR3RM INT,
cFR3SM INT)" );
#create SQL table for duplicate counts
$dbh -> do( "CREATE TABLE Duplicates (
batch INT NOT NULL,
```

```
		sample INT NOT NULL,
		ID TEXT NOT NULL,
		count INT NOT NULL)" );


		sleep(5);

		#Find input files
		my $errorfiles = "";
		my $dir = "TSVs/$f";
		chdir($dir) or die("Error: can't open $dir folder!");
		my @files = <*>; #store all files in folder
		my $temp1 = scalar(@files);
		print "$temp1 files read.\n";

		for ( my $j = 0; $j < scalar(@files); $j++ ) { #iterate over each file
			my %counttracker;
			my $filename = "$files[$j]";
			my $sample = substr($filename,0,2);
			print "Batch $f, Sample $sample\n";
			open(FILE, $filename) or die("Error: can't open $filename!");
			my @lines = <FILE>; #save file data to array of lines
			close FILE;

			if ( substr($filename,2,3) =~ /dup/ ) {
				my $dupcount = 0;
				for ( my $k = 2; $k < scalar(@lines); $k++ ) {
					$dupcount++;
					my @line = split('\t', $lines[$k]); #split line into an
array of tabs
					chomp @line;

					$dbh -> do( "INSERT INTO Duplicates VALUES (
					'$f',
					'$sample',
					'$line[0]',
					'$line[2]')" );

					# print "$line[0]\n";
				}
				if ( exists $counttracker{$sample} ) {
					if ( $counttracker{$sample} != $dupcount ) {
						print ERR "Sample $sample is
incomplete.\n";
					}
				}
```

```
                        else {
                                $counttracker{$sample} = $dupcount;
                        }
                }

                else {
                        my $tsvcount = 0;
                        for ( my $k = 1; $k < scalar(@lines); $k++ ) {
                                $tsvcount++;
                                my @line = split('\t', $lines[$k]); #split line into an
array of tabs
                                chomp @line;
                                my $temp2 = $line[82]; #Get RM AA info
                                my $temp3 = $line[84]; #Get SM AA info
                                my $temp4 = $line[81]; #Get RM nucl info
                                my $temp5 = $line[83]; #Get SM nucl info
                                #cleanup all non-essential symbols
                                $temp2 =~ tr/[]"' //d;
                                $temp3 =~ tr/[]"' //d;
                                $temp4 =~ tr/[]"' //d;
                                $temp5 =~ tr/[]"' //d;

                                my @FR3muts = pullFR3($temp4,$temp5);
#returns (mismatch nucleotides,RM count,SM count)

                                $dbh -> do( "INSERT INTO Sequence VALUES (
                                '$f',
                                '$sample',
                                '$line[0]',
                                '$line[5]',
                                '$line[6]',
                                '$line[7]',
                                '$line[8]',
                                '$line[9]',
                                '$line[10]',
                                '$line[1]',
                                '$line[2]',
                                '$line[3]',
                                '$line[11]',
                                '$line[12]',
                                '$temp2',
                                '$temp3',
                                '$temp4',
                                '$temp5',
                                '$line[51]',
                                '$line[52]',
```

```
                                        '$line[53]',
                                        '$line[54]',
                                        '$line[57]',
                                        '$line[58]',
                                        '$line[59]',
                                        '$line[60]',
                                        '$line[63]',
                                        '$line[64]',
                                        '$line[65]',
                                        '$line[66]',
                                        '$line[69]',
                                        '$line[70]',
                                        '$line[71]',
                                        '$line[72]',
                                        '$FR3muts[0]',
                                        '$FR3muts[1]',
                                        '$FR3muts[2]')" );
                                        # print "$line[13]\n";
                                }
                                if ( exists $counttracker{$sample} ) {
                                        if ( $counttracker{$sample} != $tsvcount ) {
                                                print ERR "Sample $sample is
incomplete.\n";
                                        }
                                }
                                else {
                                        $counttracker{$sample} = $tsvcount;
                                }
                        }
                }
                close ERR;
                $dbh->disconnect();
                print "Finished $database!\n";
                exit 0;
        }
        else {
                die "couldnt fork: $!\n";
        }
}

foreach (@childs) {
        my $temp = waitpid($_, 0);
        print "Done with pid $temp\n";
}

print "End of main program\n";
```

```
#####################################################
#  subroutines
#####################################################

sub pullFR3{
	my $RM = $_[0]; #argument 1 of the reference
	my $SM = $_[1]; #argument 2 of the reference
	my @FR3muts = (0,0,0); #(mismatch nucleotides,RM count,SM count)
	my @RMarray = split(',', $RM);
	my @SMarray = split(',', $SM);

	# Update mutation counts for all SM
	for ( my $s = 0; $s < scalar(@SMarray); $s++ ) {
		my $numberl = length($SMarray[$s])-6; #calculate position number
length
		my @number =
(substr($SMarray[$s],0,3),substr($SMarray[$s],3,2),substr($SMarray[$s],3+$numberl,3))
; #array = (from,position,to); don't keep position letter when present because doesn't
affect region determination
		# check if in FR3
		if ( ($number[1] > 65) && ($number[1] < 93) ) {
			$FR3muts[2] = $FR3muts[2] + 1;
			my $nuclmut = 0;
			if ( substr($number[0],0,1) ne substr($number[2],0,1) ) {
$nuclmut++; }
			if ( substr($number[0],1,1) ne substr($number[2],1,1) ) {
$nuclmut++; }
			if ( substr($number[0],2,1) ne substr($number[2],2,1) ) {
$nuclmut++; }
			if ( $nuclmut == 0 ) { print "Bad SM entry error!\n"; }
			$FR3muts[0] = $FR3muts[0] + $nuclmut;
		}
	}
	# Update mutation counts for all RM
	for ( my $r = 0; $r < scalar(@RMarray); $r++ ) {
		my $numberl = length($RMarray[$r])-6; #calculate position number
length
		my @number =
(substr($RMarray[$r],0,3),substr($RMarray[$r],3,2),substr($RMarray[$r],3+$numberl,3)
); #array = (from,position,to); don't keep position letter when present because doesn't
affect region determination
		# check if in FR3
		if ( ($number[1] > 65) && ($number[1] < 93) ) {
			$FR3muts[1] = $FR3muts[1] + 1;
			my $nuclmut = 0;
```

```
                        if ( substr($number[0],0,1) ne substr($number[2],0,1) ) {
$nuclmut++; }
                        if ( substr($number[0],1,1) ne substr($number[2],1,1) ) {
$nuclmut++; }
                        if ( substr($number[0],2,1) ne substr($number[2],2,1) ) {
$nuclmut++; }
                        if ( $nuclmut == 0 ) { print "Bad RM entry error!\n"; }
                        $FR3muts[0] = $FR3muts[0] + $nuclmut;
                }
        }

        return @FR3muts;
}
```

APPENDIX 2

SQLite script to filter out sequences from database.

Run on the SequenceRAW SQLite database of aligned sequence information and matching sequence count information. Also uses a user-inputted table called "diagnosisinput", which can be added using a tab delimited table with the following categories: "batch" (number); "sample" (number); "Diogenix" (patient identifier); "diagnosis" (number code to be set by the user; in the Validation study, 1 = $RRMS_{3/3}$, 2 = $RRMS_{2/3}$, 3 = $OND_{3/3}$, 4 = $OND_{2/3}$); "Dremove" (diagnosis-based removal code, used to mark with a "1" any sample that is excluded regardless of diagnosis code, such as ONDs with possible MS or NMO samples; samples not to be excluded get a "0"). SQLite step descriptions are preceded with "--". Filtering steps are in the exact order presented in Figure 2-2 which provides a sequence number summary of each filter step impact on sequence counts.

```
-- ValidationClean6.sql
CREATE TABLE Sequence
AS SELECT SequenceRAW.*, diagnosisinput.Diogenix, diagnosisinput.diagnosis,
diagnosisinput.Dremove
FROM SequenceRAW
LEFT JOIN diagnosisinput
ON (SequenceRAW.batch = diagnosisinput.batch AND SequenceRAW.sample =
diagnosisinput.sample);

CREATE TABLE OUTraw
AS SELECT Sequence.batch AS batch, Sequence.sample AS sample,
SUM(Sequence.count) AS count
FROM Sequence
GROUP BY Sequence.batch, Sequence.sample;

--not VH4 filter count
CREATE TABLE OUTnotvh4
```

```
AS SELECT Sequence.batch AS batch, Sequence.sample AS sample,
SUM(Sequence.count) AS count
FROM Sequence
WHERE Sequence.vgene NOT LIKE 'IGHV4%'
GROUP BY Sequence.batch, Sequence.sample;

--N nucleotides in the sequence filter count
CREATE TABLE OUTns
AS SELECT Sequence.batch AS batch, Sequence.sample AS sample,
SUM(Sequence.count) AS count
FROM Sequence
WHERE Sequence.agsRM LIKE '%X%' OR Sequence.agsSM LIKE '%X%' OR
(Sequence.cdr3aa != "" AND Sequence.cdr3 LIKE '%N%')
GROUP BY Sequence.batch, Sequence.sample;

--All filter criteria 1 failed sequences removed
CREATE TABLE Sequencefiltered
AS SELECT Sequence.*,
SUBSTR(Sequence.vgene,1,LENGTH(REPLACE(Sequence.vgene,'*',''))-2) AS vhgene,
SUBSTR(Sequence.jgene,1,LENGTH(REPLACE(Sequence.jgene,'*',''))-2) AS jhgene
FROM Sequence
WHERE Sequence.vgene LIKE 'IGHV4%' AND Sequence.vgene NOT LIKE '%OR15-
8%' AND Sequence.vgene NOT LIKE 'IGHV4-55%'
AND Sequence.outofframe = "False" AND Sequence.missCYS = "False" AND
Sequence.missTRP = "False" AND Sequence.stopcodon = "False" AND Sequence.indels
= "False"
AND Sequence.cdr3aa != "" AND Sequence.cdr3 NOT LIKE '%N%' AND
Sequence.agsRM NOT LIKE '%X%' AND Sequence.agsSM NOT LIKE '%X%';

--Truncated CDR or FR regions filter
CREATE TABLE Sequencefiltered2
AS SELECT Sequencefiltered.*
FROM Sequencefiltered
WHERE (Sequencefiltered.FR2l = 42 AND Sequencefiltered.FR3l > 89)
AND (
((Sequencefiltered.vhgene LIKE 'IGHV4-30%' OR Sequencefiltered.vhgene LIKE
'IGHV4-31%' OR Sequencefiltered.vhgene LIKE 'IGHV4-39%' OR
Sequencefiltered.vhgene LIKE 'IGHV4-61%') AND (Sequencefiltered.CDR1l = 21))
OR ((Sequencefiltered.vhgene LIKE 'IGHV4-28%' OR Sequencefiltered.vhgene LIKE
'IGHV4-38%' OR Sequencefiltered.vhgene LIKE 'IGHV4-4%') AND
(Sequencefiltered.vgene NOT LIKE 'IGHV4-4*07' AND Sequencefiltered.vgene NOT
LIKE 'IGHV4-4*08' AND Sequencefiltered.CDR1l = 18))
OR ((Sequencefiltered.vhgene LIKE 'IGHV4-34%' OR Sequencefiltered.vhgene LIKE
'IGHV4-59%' OR Sequencefiltered.vgene LIKE 'IGHV4-4*07' OR
Sequencefiltered.vgene LIKE 'IGHV4-4*08') AND (Sequencefiltered.CDR1l = 15))
);
```

```
--Homology < 85% filter prepare
CREATE TABLE Sequencefiltered3
AS SELECT Sequencefiltered2.*, 1-
SUM(Sequencefiltered2.CDR1mut+Sequencefiltered2.FR2mut+Sequencefiltered2.CDR2
mut+Sequencefiltered2.cFR3mut)*1.0/SUM(Sequencefiltered2.CDR1l+Sequencefiltered
2.FR2l+Sequencefiltered2.CDR2l+90) AS homology
FROM Sequencefiltered2
GROUP BY Sequencefiltered2.ID;

--Homology < 85% filter
CREATE TABLE Sequencefiltered4
AS SELECT Sequencefiltered3.*, 90 AS cFR3l
FROM Sequencefiltered3
WHERE Sequencefiltered3.homology > 0.85;

-- check vhdiversity
SELECT Sequencefiltered4.vhgene AS vhgene, COUNT (Sequencefiltered4.ID) AS
count
FROM Sequencefiltered4
GROUP BY Sequencefiltered4.vhgene;

CREATE TABLE OUTfiltered
AS SELECT Sequencefiltered4.batch AS batch, Sequencefiltered4.sample AS sample,
SUM(Sequencefiltered4.count) AS count
FROM Sequencefiltered4
GROUP BY Sequencefiltered4.batch, Sequencefiltered4.sample;

-- Unique sequence defined here
CREATE TABLE SequencefilteredUNIQUE
AS SELECT Sequencefiltered4.*, SUM(Sequencefiltered4.count) AS finalcount,
Sequencefiltered4.Diogenix AS cDiogenix
FROM Sequencefiltered4
WHERE Sequencefiltered4.Diogenix NOT LIKE "M106" AND
Sequencefiltered4.Diogenix NOT LIKE "N106" AND Sequencefiltered4.Diogenix NOT
LIKE "M105" AND Sequencefiltered4.Diogenix NOT LIKE "N105"
GROUP BY Sequencefiltered4.batch, Sequencefiltered4.sample,
Sequencefiltered4.vhgene, Sequencefiltered4.jhgene, Sequencefiltered4.cdr3,
Sequencefiltered4.agsRM;
INSERT INTO SequencefilteredUNIQUE
SELECT Sequencefiltered4.*, SUM(Sequencefiltered4.count) AS finalcount, 106 AS
cDiogenix
FROM Sequencefiltered4
WHERE Sequencefiltered4.Diogenix LIKE "M106" OR Sequencefiltered4.Diogenix
LIKE "N106"
```

```
GROUP BY Sequencefiltered4.batch, Sequencefiltered4.sample,
Sequencefiltered4.vhgene, Sequencefiltered4.jhgene, Sequencefiltered4.cdr3,
Sequencefiltered4.agsRM;
INSERT INTO SequencefilteredUNIQUE
SELECT Sequencefiltered4.*, SUM(Sequencefiltered4.count) AS finalcount, 105 AS
cDiogenix
FROM Sequencefiltered4
WHERE Sequencefiltered4.Diogenix LIKE "M105" OR Sequencefiltered4.Diogenix
LIKE "N105"
GROUP BY Sequencefiltered4.batch, Sequencefiltered4.sample,
Sequencefiltered4.vhgene, Sequencefiltered4.jhgene, Sequencefiltered4.cdr3,
Sequencefiltered4.agsRM;

-- Prepare for Crossover sequence filtering
CREATE TABLE CDR3list
AS SELECT SequencefilteredUNIQUE.cdr3 AS cdr3,
SequencefilteredUNIQUE.cDiogenix AS Diogenix,
SUM(SequencefilteredUNIQUE.finalcount) AS count
FROM SequencefilteredUNIQUE
GROUP BY SequencefilteredUNIQUE.cdr3, SequencefilteredUNIQUE.cDiogenix;

CREATE TABLE COlist
AS SELECT CDR3list.cdr3 AS cdr3, COUNT(CDR3list.Diogenix) AS samples,
SUM(CDR3list.count) AS count
FROM CDR3list
GROUP BY CDR3list.cdr3;

CREATE TABLE CDR3math
AS SELECT CDR3list.*, (CDR3list.count * 1.0 / COlist.count) AS percentsample
FROM CDR3list, COlist
WHERE COlist.samples > 1 AND CDR3list.cdr3 = COlist.cdr3;

CREATE TABLE COblacklist
AS SELECT CDR3math.*
FROM CDR3math
WHERE CDR3math.percentsample < 0.99;

-- Crossover sequence filtering preparation
CREATE TABLE SequencefilteredCO
AS SELECT SequencefilteredUNIQUE.*, COblacklist.percentsample AS percentsample
FROM SequencefilteredUNIQUE
LEFT JOIN COblacklist
ON (SequencefilteredUNIQUE.cdr3 = COblacklist.cdr3 AND
SequencefilteredUNIQUE.cDiogenix = COblacklist.Diogenix);

-- Count R1 sequences for no CO sequences
```

```
CREATE TABLE OUTfilteredR1
AS SELECT SequencefilteredCO.batch AS batch, SequencefilteredCO.sample AS
sample, SUM(SequencefilteredCO.finalcount) AS count
FROM SequencefilteredCO
WHERE SequencefilteredCO.finalcount > 1
GROUP BY SequencefilteredCO.batch, SequencefilteredCO.sample;

-- Count R1 sequences for CO sequences
CREATE TABLE OUTfilteredR1CO
AS SELECT SequencefilteredCO.batch AS batch, SequencefilteredCO.sample AS
sample, SUM(SequencefilteredCO.finalcount) AS count
FROM SequencefilteredCO
WHERE SequencefilteredCO.percentsample IS NOT NULL AND
SequencefilteredCO.finalcount > 1
GROUP BY SequencefilteredCO.batch, SequencefilteredCO.sample;

-- Crossover sequence filtering
CREATE TABLE SequencefilteredCORE
AS SELECT SequencefilteredCO.*
FROM SequencefilteredCO
WHERE SequencefilteredCO.percentsample IS NULL;

-- R0 sequence filtering
CREATE TABLE SequencefilteredR1
AS SELECT SequencefilteredCORE.*,
SequencefilteredCORE.CDR1RM+SequencefilteredCORE.FR2RM+SequencefilteredCO
RE.CDR2RM+SequencefilteredCORE.cFR3RM AS RMcount
FROM SequencefilteredCORE
WHERE SequencefilteredCORE.finalcount > 1;

-- Prepare for AGS calculation
CREATE TABLE SequencefilteredR1AGS31B
AS SELECT SequencefilteredR1.ID AS ID, COUNT(SequencefilteredR1.ID) AS count
FROM SequencefilteredR1
WHERE SequencefilteredR1.agsRM LIKE '%31B%'
GROUP BY SequencefilteredR1.ID;
CREATE TABLE SequencefilteredR1AGS40
AS SELECT SequencefilteredR1.ID AS ID, COUNT(SequencefilteredR1.ID) AS count
FROM SequencefilteredR1
WHERE SequencefilteredR1.agsRM LIKE '%40%'
GROUP BY SequencefilteredR1.ID;
CREATE TABLE SequencefilteredR1AGS56
AS SELECT SequencefilteredR1.ID AS ID, COUNT(SequencefilteredR1.ID) AS count
FROM SequencefilteredR1
WHERE SequencefilteredR1.agsRM LIKE '%56%'
GROUP BY SequencefilteredR1.ID;
```

```
CREATE TABLE SequencefilteredR1AGS57
AS SELECT SequencefilteredR1.ID AS ID, COUNT(SequencefilteredR1.ID) AS count
FROM SequencefilteredR1
WHERE SequencefilteredR1.agsRM LIKE '%57%'
GROUP BY SequencefilteredR1.ID;
CREATE TABLE SequencefilteredR1AGS81
AS SELECT SequencefilteredR1.ID AS ID, COUNT(SequencefilteredR1.ID) AS count
FROM SequencefilteredR1
WHERE SequencefilteredR1.agsRM LIKE '%81%'
GROUP BY SequencefilteredR1.ID;
CREATE TABLE SequencefilteredR1AGS89
AS SELECT SequencefilteredR1.ID AS ID, COUNT(SequencefilteredR1.ID) AS count
FROM SequencefilteredR1
WHERE SequencefilteredR1.agsRM LIKE '%89%'
GROUP BY SequencefilteredR1.ID;

CREATE TABLE SequencefilteredR1RM
AS SELECT SequencefilteredR1.*, COALESCE(SequencefilteredR1AGS31B.count,0)
AS AGS31B, COALESCE(SequencefilteredR1AGS40.count,0) AS AGS40,
COALESCE(SequencefilteredR1AGS56.count,0) AS AGS56,
COALESCE(SequencefilteredR1AGS57.count,0) AS AGS57,
COALESCE(SequencefilteredR1AGS81.count,0) AS AGS81,
COALESCE(SequencefilteredR1AGS89.count,0) AS AGS89
FROM SequencefilteredR1
LEFT JOIN SequencefilteredR1AGS31B
ON (SequencefilteredR1.ID = SequencefilteredR1AGS31B.ID)
LEFT JOIN SequencefilteredR1AGS40
ON (SequencefilteredR1.ID = SequencefilteredR1AGS40.ID)
LEFT JOIN SequencefilteredR1AGS56
ON (SequencefilteredR1.ID = SequencefilteredR1AGS56.ID)
LEFT JOIN SequencefilteredR1AGS57
ON (SequencefilteredR1.ID = SequencefilteredR1AGS57.ID)
LEFT JOIN SequencefilteredR1AGS81
ON (SequencefilteredR1.ID = SequencefilteredR1AGS81.ID)
LEFT JOIN SequencefilteredR1AGS89
ON (SequencefilteredR1.ID = SequencefilteredR1AGS89.ID);

CREATE TABLE OUTuniqueClean
AS SELECT SequencefilteredR1RM.batch AS batch, SequencefilteredR1RM.sample AS
sample, COUNT(SequencefilteredR1RM.ID) AS count
FROM SequencefilteredR1RM
GROUP BY SequencefilteredR1RM.batch, SequencefilteredR1RM.sample;

CREATE TABLE SequencefilteredR1RMfinal
AS SELECT SequencefilteredR1RM.*, SUBSTR(SequencefilteredR1RM.vhgene,1,8)
AS vhgeneclean
```

```
FROM SequencefilteredR1RM;

-- 8 VH4 genes kept for Diversity Index calculations
CREATE TABLE SequencefilteredR1RMfinalDI
AS SELECT SequencefilteredR1RM.*, SUBSTR(SequencefilteredR1RM.vhgene,1,8)
AS vhgeneclean
FROM SequencefilteredR1RM
WHERE SequencefilteredR1RM.vhgene NOT LIKE 'IGHV4-28';

-- check vhdiversity
SELECT SequencefilteredR1RMfinal.vhgeneclean AS vhgeneclean, COUNT
(SequencefilteredR1RMfinal.ID) AS count
FROM SequencefilteredR1RMfinal
GROUP BY SequencefilteredR1RMfinal.vhgeneclean;

CREATE TABLE OUTuniqueCleanNOORF
AS SELECT SequencefilteredR1RMfinal.batch AS batch,
SequencefilteredR1RMfinal.sample AS sample,
COUNT(SequencefilteredR1RMfinal.ID) AS count
FROM SequencefilteredR1RMfinal
GROUP BY SequencefilteredR1RMfinal.batch, SequencefilteredR1RMfinal.sample;

CREATE TABLE OUTuniqueCleanNOORFDI
AS SELECT SequencefilteredR1RMfinalDI.batch AS batch,
SequencefilteredR1RMfinalDI.sample AS sample,
COUNT(SequencefilteredR1RMfinalDI.ID) AS count
FROM SequencefilteredR1RMfinalDI
GROUP BY SequencefilteredR1RMfinalDI.batch,
SequencefilteredR1RMfinalDI.sample;

CREATE TABLE OUTdiversityindex
AS SELECT SequencefilteredR1RMfinalDI.batch AS batch,
SequencefilteredR1RMfinalDI.sample AS sample,
SequencefilteredR1RMfinalDI.vhgeneclean AS vhgeneclean,
COUNT(SequencefilteredR1RMfinalDI.ID) * 1.0 / OUTuniqueCleanNOORFDI.count
AS fraction
FROM SequencefilteredR1RMfinalDI
LEFT JOIN OUTuniqueCleanNOORFDI
ON (SequencefilteredR1RMfinalDI.batch = OUTuniqueCleanNOORFDI.batch AND
SequencefilteredR1RMfinalDI.sample = OUTuniqueCleanNOORFDI.sample)
GROUP BY SequencefilteredR1RMfinalDI.batch,
SequencefilteredR1RMfinalDI.sample, SequencefilteredR1RMfinalDI.vhgeneclean;

-- Get MSPrecise output
SELECT SequencefilteredR1RMfinal.batch, SequencefilteredR1RMfinal.sample,
SequencefilteredR1RMfinal.Diogenix, COUNT(SequencefilteredR1RMfinal.ID) AS
```

```
count, SequencefilteredR1RMfinal.diagnosis,
SUM(SequencefilteredR1RMfinal.RMcount) AS RMcount,
SUM(SequencefilteredR1RMfinal.AGS31B+SequencefilteredR1RMfinal.AGS40+Seque
ncefilteredR1RMfinal.AGS56+SequencefilteredR1RMfinal.AGS57+SequencefilteredR1
RMfinal.AGS81) AS AGS5rm
FROM SequencefilteredR1RMfinal
GROUP BY SequencefilteredR1RMfinal.batch, SequencefilteredR1RMfinal.sample;

-- Find dominant seq difference
CREATE TABLE SeqCountMAX
AS SELECT SequencefilteredR1RMfinal.batch AS batch,
SequencefilteredR1RMfinal.sample AS sample, SequencefilteredR1RMfinal.Diogenix
AS Diogenix, MAX(SequencefilteredR1RMfinal.finalcount) AS MAXcount,
SequencefilteredR1RMfinal.ID AS ID
FROM SequencefilteredR1RMfinal
GROUP BY SequencefilteredR1RMfinal.batch, SequencefilteredR1RMfinal.sample;

CREATE TABLE SeqCountMAX2
AS SELECT SequencefilteredR1RMfinal.*
FROM SequencefilteredR1RMfinal
LEFT JOIN SeqCountMAX
ON (SequencefilteredR1RMfinal.batch = SeqCountMAX.batch AND
SequencefilteredR1RMfinal.sample = SeqCountMAX.sample)
WHERE SequencefilteredR1RMfinal.ID != SeqCountMAX.ID;

CREATE TABLE SeqCountMAX3
AS SELECT SeqCountMAX2.batch AS batch, SeqCountMAX2.sample AS sample,
SeqCountMAX2.Diogenix AS Diogenix, MAX(SeqCountMAX2.finalcount) AS
MAXcount, SeqCountMAX2.ID AS ID
FROM SeqCountMAX2
GROUP BY SeqCountMAX2.batch, SeqCountMAX2.sample;

CREATE TABLE SeqCountMAX4
AS SELECT SeqCountMAX.batch AS batch, SeqCountMAX.sample AS sample,
SeqCountMAX.Diogenix AS Diogenix, SeqCountMAX3.MAXcount * 1.0 /
SeqCountMAX.MAXcount AS High2high
FROM SeqCountMAX
LEFT JOIN SeqCountMAX3
ON (SeqCountMAX.batch = SeqCountMAX3.batch AND SeqCountMAX.sample =
SeqCountMAX3.sample);

-- Generate final output table for analysis of filter counts
CREATE TABLE OUTFINAL
AS SELECT OUTraw.batch AS batch, OUTraw.sample AS sample,
diagnosisinput.Diogenix AS Diogenix, diagnosisinput.diagnosis AS diagnosis,
diagnosisinput.Dremove AS Dremove, OUTraw.count AS raw,
```

```
COALESCE(OUTnotvh4.count,0) AS notvh4, COALESCE(OUTfiltered.count,0) AS
filtered, COALESCE(OUTfilteredR1.count,0) AS filteredR1,
COALESCE(OUTfilteredR1CO.count,0) AS filteredR1CO,
COALESCE(OUTuniqueClean.count,0) AS uniqueClean,
COALESCE(OUTuniqueCleanNOORF.count,0) AS uniqueCleanNOORF,
COALESCE(SeqCountMAX4.High2high,0) AS High2high
FROM OUTraw
LEFT JOIN OUTnotvh4
ON (OUTraw.batch = OUTnotvh4.batch AND OUTraw.sample = OUTnotvh4.sample)
LEFT JOIN OUTfiltered
ON (OUTraw.batch = OUTfiltered.batch AND OUTraw.sample = OUTfiltered.sample)
LEFT JOIN OUTfilteredR1
ON (OUTraw.batch = OUTfilteredR1.batch AND OUTraw.sample =
OUTfilteredR1.sample)
LEFT JOIN OUTfilteredR1CO
ON (OUTraw.batch = OUTfilteredR1CO.batch AND OUTraw.sample =
OUTfilteredR1CO.sample)
LEFT JOIN OUTuniqueClean
ON (OUTraw.batch = OUTuniqueClean.batch AND OUTraw.sample =
OUTuniqueClean.sample)
LEFT JOIN OUTuniqueCleanNOORF
ON (OUTraw.batch = OUTuniqueCleanNOORF.batch AND OUTraw.sample =
OUTuniqueCleanNOORF.sample)
LEFT JOIN diagnosisinput
ON (OUTraw.batch = diagnosisinput.batch AND OUTraw.sample =
diagnosisinput.sample)
LEFT JOIN SeqCountMAX4
ON (OUTraw.batch = SeqCountMAX4.batch AND OUTraw.sample =
SeqCountMAX4.sample);
```

APPENDIX 3

SQLite script to filter out samples from database and generate final sample level genetic analysis data.

Runs directly on the tables generated by the cleaning script (Appendix 2). Data tables that are specifically for output are indicated by an "OUT" in the name. SQLite step descriptions are preceded with "--".

```
-- ValidationAnalyze4.sql
CREATE TABLE AnalyzeSamplesDetailed1
AS SELECT OUTFINAL.*, OUTFINAL.notvh4 * 1.0 / OUTFINAL.raw AS pVH4,
COALESCE(OUTFINAL.filteredR1CO * 1.0 / OUTFINAL.filteredR1,0) AS pCO
FROM OUTFINAL
GROUP BY OUTFINAL.batch, OUTFINAL.sample;

-- identify duplicates
CREATE TABLE AnalyzeDup1
AS SELECT AnalyzeSamplesDetailed1.Diogenix AS Diogenix,
COUNT(AnalyzeSamplesDetailed1.Diogenix) as Diogenixcount
FROM AnalyzeSamplesDetailed1
GROUP BY AnalyzeSamplesDetailed1.Diogenix;

CREATE TABLE AnalyzeSamplesDetailed2
AS SELECT AnalyzeSamplesDetailed1.*, AnalyzeDup1.Diogenixcount AS dup
FROM AnalyzeSamplesDetailed1
LEFT JOIN AnalyzeDup1
ON (AnalyzeSamplesDetailed1.Diogenix = AnalyzeDup1.Diogenix);

CREATE TABLE AnalyzeDup2
AS SELECT AnalyzeDup1.*
FROM AnalyzeDup1
WHERE AnalyzeDup1.Diogenixcount > 1;

CREATE TABLE AnalyzeDup3
AS SELECT AnalyzeSamplesDetailed1.Diogenix AS Diogenix,
AnalyzeSamplesDetailed1.uniqueCleanNOORF AS uniqueCleanNOORF
FROM AnalyzeSamplesDetailed1
INNER JOIN AnalyzeDup2
ON (AnalyzeSamplesDetailed1.Diogenix = AnalyzeDup2.Diogenix);
```

```
CREATE TABLE AnalyzeDup4
AS SELECT AnalyzeDup3.*, MAX(AnalyzeDup3.uniqueCleanNOORF) AS Maxseq
FROM AnalyzeDup3
GROUP BY AnalyzeDup3.Diogenix;

-- remove duplicates
CREATE TABLE AnalyzeSamplesDetailed
AS SELECT AnalyzeSamplesDetailed2.*
FROM AnalyzeSamplesDetailed2
LEFT JOIN AnalyzeDup4
ON (AnalyzeSamplesDetailed2.Diogenix = AnalyzeDup4.Diogenix)
WHERE AnalyzeSamplesDetailed2.dup = 1 OR
AnalyzeSamplesDetailed2.uniqueCleanNOORF = AnalyzeDup4.Maxseq;

-- Genetics: 3/3 and 2/3 without lowseq
CREATE TABLE AnalyzeRemove1
AS SELECT AnalyzeSamplesDetailed.*
FROM AnalyzeSamplesDetailed
WHERE AnalyzeSamplesDetailed.Dremove = 0 AND
(AnalyzeSamplesDetailed.diagnosis = 1 OR AnalyzeSamplesDetailed.diagnosis = 3 OR
AnalyzeSamplesDetailed.diagnosis = 2 OR AnalyzeSamplesDetailed.diagnosis = 4)
AND AnalyzeSamplesDetailed.pCO < 0.5
AND (AnalyzeSamplesDetailed.uniqueCleanNOORF > 24 OR
(AnalyzeSamplesDetailed.High2high > 0.02 AND
AnalyzeSamplesDetailed.uniqueCleanNOORF > 8));

-- Genetics: 3/3 and 2/3 only lowseq
CREATE TABLE AnalyzeRemove4
AS SELECT AnalyzeSamplesDetailed.*
FROM AnalyzeSamplesDetailed
WHERE AnalyzeSamplesDetailed.Dremove = 0 AND
(AnalyzeSamplesDetailed.diagnosis = 1 OR AnalyzeSamplesDetailed.diagnosis = 3 OR
AnalyzeSamplesDetailed.diagnosis = 2 OR AnalyzeSamplesDetailed.diagnosis = 4)
AND AnalyzeSamplesDetailed.pCO < 0.5
AND (AnalyzeSamplesDetailed.uniqueCleanNOORF < 9 OR
(AnalyzeSamplesDetailed.High2high < 0.02 AND
AnalyzeSamplesDetailed.uniqueCleanNOORF < 25));

-- MSPrecise: 3/3 without lowseq
CREATE TABLE AnalyzeRemove2
AS SELECT AnalyzeSamplesDetailed.*
FROM AnalyzeSamplesDetailed
WHERE AnalyzeSamplesDetailed.Dremove = 0 AND
(AnalyzeSamplesDetailed.diagnosis = 1 OR AnalyzeSamplesDetailed.diagnosis = 3)
AND AnalyzeSamplesDetailed.pCO < 0.5
```

```
AND (AnalyzeSamplesDetailed.uniqueCleanNOORF > 24 OR
(AnalyzeSamplesDetailed.High2high > 0.02 AND
AnalyzeSamplesDetailed.uniqueCleanNOORF > 8));

-- MSPrecise: 3/3 only lowseq
CREATE TABLE AnalyzeRemove3
AS SELECT AnalyzeSamplesDetailed.*
FROM AnalyzeSamplesDetailed
WHERE AnalyzeSamplesDetailed.Dremove = 0 AND
(AnalyzeSamplesDetailed.diagnosis = 1 OR AnalyzeSamplesDetailed.diagnosis = 3)
AND AnalyzeSamplesDetailed.pCO < 0.5
AND (AnalyzeSamplesDetailed.uniqueCleanNOORF < 9 OR
(AnalyzeSamplesDetailed.High2high < 0.02 AND
AnalyzeSamplesDetailed.uniqueCleanNOORF < 25));



-- Generate genetics repertoire outputs
CREATE TABLE AnalyzeSequence
AS SELECT SequencefilteredR1RMfinal.*
FROM AnalyzeRemove1
LEFT JOIN SequencefilteredR1RMfinal
ON (AnalyzeRemove1.batch = SequencefilteredR1RMfinal.batch AND
AnalyzeRemove1.sample = SequencefilteredR1RMfinal.sample);

CREATE TABLE AnalyzeOUTvh4
AS SELECT AnalyzeSequence.diagnosis AS diagnosis, AnalyzeSequence.Diogenix AS
Diogenix, AnalyzeSequence.vhgeneclean AS vhgene, COUNT(AnalyzeSequence.ID) AS
count
FROM AnalyzeSequence
GROUP BY AnalyzeSequence.Diogenix, AnalyzeSequence.vhgeneclean
ORDER BY AnalyzeSequence.diagnosis, AnalyzeSequence.Diogenix;

CREATE TABLE AnalyzeOUTjh
AS SELECT AnalyzeSequence.diagnosis AS diagnosis, AnalyzeSequence.Diogenix AS
Diogenix, AnalyzeSequence.jhgene AS jhgene, COUNT(AnalyzeSequence.ID) AS count
FROM AnalyzeSequence
GROUP BY AnalyzeSequence.Diogenix, AnalyzeSequence.jhgene
ORDER BY AnalyzeSequence.diagnosis, AnalyzeSequence.Diogenix;

CREATE TABLE AnalyzeOUTmf
AS SELECT AnalyzeSequence.diagnosis AS diagnosis, AnalyzeSequence.Diogenix AS
Diogenix, SUM(AnalyzeSequence.CDR1mut+AnalyzeSequence.CDR2mut) AS
CDRmut, SUM(AnalyzeSequence.FR2mut+AnalyzeSequence.cFR3mut) AS FRmut,
SUM(AnalyzeSequence.CDR1l+AnalyzeSequence.CDR2l) AS CDRl,
SUM(AnalyzeSequence.FR2l+AnalyzeSequence.cFR3l) AS FRl
```

```
FROM AnalyzeSequence
GROUP BY AnalyzeSequence.Diogenix
ORDER BY AnalyzeSequence.diagnosis, AnalyzeSequence.Diogenix;

CREATE TABLE AnalyzeOUTrmf
AS SELECT AnalyzeSequence.diagnosis AS diagnosis, AnalyzeSequence.Diogenix AS
Diogenix, SUM(AnalyzeSequence.CDR1RM+AnalyzeSequence.CDR2RM) AS CDRrm,
SUM(AnalyzeSequence.FR2RM+AnalyzeSequence.cFR3RM) AS FRrm,
SUM(AnalyzeSequence.CDR1l+AnalyzeSequence.CDR2l) AS CDRl,
SUM(AnalyzeSequence.FR2l+AnalyzeSequence.cFR3l) AS FRl,
SUM(AnalyzeSequence.CDR1SM+AnalyzeSequence.CDR2SM) AS CDRsm,
SUM(AnalyzeSequence.FR2SM+AnalyzeSequence.cFR3SM) AS FRsm
FROM AnalyzeSequence
GROUP BY AnalyzeSequence.Diogenix
ORDER BY AnalyzeSequence.diagnosis, AnalyzeSequence.Diogenix;

CREATE TABLE AnalyzeOUTcdr3l
AS SELECT AnalyzeSequence.diagnosis AS diagnosis, AnalyzeSequence.Diogenix AS
Diogenix, SUM(LENGTH(AnalyzeSequence.cdr3aa))*1.0 /
COUNT(AnalyzeSequence.ID) AS cdr3l
FROM AnalyzeSequence
GROUP BY AnalyzeSequence.Diogenix
ORDER BY AnalyzeSequence.diagnosis, AnalyzeSequence.Diogenix;

CREATE TABLE Analyzecdr3charge
AS SELECT AnalyzeSequence.diagnosis AS diagnosis, AnalyzeSequence.Diogenix AS
Diogenix, LENGTH(AnalyzeSequence.cdr3aa) -
LENGTH(REPLACE(AnalyzeSequence.cdr3aa,'R','')) AS R,
LENGTH(AnalyzeSequence.cdr3aa) -
LENGTH(REPLACE(AnalyzeSequence.cdr3aa,'K','')) AS K,
LENGTH(AnalyzeSequence.cdr3aa) -
LENGTH(REPLACE(AnalyzeSequence.cdr3aa,'D','')) AS D,
LENGTH(AnalyzeSequence.cdr3aa) -
LENGTH(REPLACE(AnalyzeSequence.cdr3aa,'E','')) AS E
FROM AnalyzeSequence;

CREATE TABLE AnalyzeOUTcdr3charge
AS SELECT Analyzecdr3charge.diagnosis AS diagnosis, Analyzecdr3charge.Diogenix
AS Diogenix, SUM(Analyzecdr3charge.R+Analyzecdr3charge.K-Analyzecdr3charge.D-
Analyzecdr3charge.E)*1.0 / COUNT(Analyzecdr3charge.Diogenix) AS cdr3charge
FROM Analyzecdr3charge
GROUP BY Analyzecdr3charge.Diogenix
ORDER BY Analyzecdr3charge.diagnosis, Analyzecdr3charge.Diogenix;

CREATE TABLE AnalyzedClone1
```

```
AS SELECT AnalyzeSequence.diagnosis AS diagnosis, AnalyzeSequence.Diogenix AS
Diogenix, AnalyzeSequence.cdr3 AS cdr3, COUNT(AnalyzeSequence.ID) AS count
FROM AnalyzeSequence
GROUP BY AnalyzeSequence.Diogenix, AnalyzeSequence.cdr3;

CREATE TABLE AnalyzedClone2
AS SELECT AnalyzedClone1.diagnosis AS diagnosis, AnalyzedClone1.Diogenix AS
Diogenix, COUNT(AnalyzedClone1.cdr3) AS clones
FROM AnalyzedClone1
WHERE AnalyzedClone1.count > 1
GROUP BY AnalyzedClone1.Diogenix;

CREATE TABLE AnalyzedClone3
AS SELECT AnalyzedClone1.diagnosis AS diagnosis, AnalyzedClone1.Diogenix AS
Diogenix, COUNT(AnalyzedClone1.cdr3) AS clonetotal
FROM AnalyzedClone1
GROUP BY AnalyzedClone1.Diogenix;

CREATE TABLE AnalyzedOUTClone
AS SELECT AnalyzedClone3.*, COALESCE(AnalyzedClone2.clones,0) AS clones
FROM AnalyzedClone3
LEFT JOIN AnalyzedClone2
ON (AnalyzedClone3.Diogenix = AnalyzedClone2.Diogenix)
ORDER BY AnalyzedClone3.diagnosis, AnalyzedClone3.Diogenix;

-- Count reads with x muts OPTION 1
CREATE TABLE Analyzed_seqmuts
AS SELECT AnalyzeSequence.diagnosis AS diagnosis, AnalyzeSequence.Diogenix AS
Diogenix,
AnalyzeSequence.CDR1mut+AnalyzeSequence.FR2mut+AnalyzeSequence.CDR2mut+
AnalyzeSequence.cFR3mut AS mutcount
FROM AnalyzeSequence;

-- Count reads with x RMs OPTION 2
CREATE TABLE Analyzed_seqmuts
AS SELECT AnalyzeSequence.diagnosis AS diagnosis, AnalyzeSequence.Diogenix AS
Diogenix,
AnalyzeSequence.CDR1RM+AnalyzeSequence.FR2RM+AnalyzeSequence.CDR2RM+
AnalyzeSequence.cFR3RM AS mutcount
FROM AnalyzeSequence;

CREATE TABLE Analyzed_seqmuts1
AS SELECT Analyzed_seqmuts.diagnosis AS diagnosis, Analyzed_seqmuts.Diogenix
AS Diogenix, COUNT(Analyzed_seqmuts.Diogenix) AS count
FROM Analyzed_seqmuts
GROUP BY Analyzed_seqmuts.Diogenix;
```

```
CREATE TABLE Analyzed_seqmuts2
AS SELECT Analyzed_seqmuts.diagnosis AS diagnosis, Analyzed_seqmuts.Diogenix
AS Diogenix, COUNT(Analyzed_seqmuts.Diogenix) AS count
FROM Analyzed_seqmuts
WHERE Analyzed_seqmuts.mutcount = 0
GROUP BY Analyzed_seqmuts.Diogenix;

CREATE TABLE Analyzed_seqmuts3
AS SELECT Analyzed_seqmuts.diagnosis AS diagnosis, Analyzed_seqmuts.Diogenix
AS Diogenix, COUNT(Analyzed_seqmuts.Diogenix) AS count
FROM Analyzed_seqmuts
WHERE Analyzed_seqmuts.mutcount = 1
GROUP BY Analyzed_seqmuts.Diogenix;

CREATE TABLE AnalyzedOUT_seqmuts
AS SELECT Analyzed_seqmuts1.*, COALESCE(Analyzed_seqmuts2.count,0)*100.0 /
Analyzed_seqmuts1.count AS p0, COALESCE(Analyzed_seqmuts3.count,0)*100.0 /
Analyzed_seqmuts1.count AS p1, (Analyzed_seqmuts1.count -
COALESCE(Analyzed_seqmuts2.count,0) -
COALESCE(Analyzed_seqmuts3.count,0))*100.0 / Analyzed_seqmuts1.count AS p2
FROM Analyzed_seqmuts1
LEFT JOIN Analyzed_seqmuts2
ON (Analyzed_seqmuts1.Diogenix = Analyzed_seqmuts2.Diogenix)
LEFT JOIN Analyzed_seqmuts3
ON (Analyzed_seqmuts1.Diogenix = Analyzed_seqmuts3.Diogenix);

CREATE TABLE AnalyzedOUT_seqmutsEXT
AS SELECT Analyzed_seqmuts.diagnosis AS diagnosis, Analyzed_seqmuts.mutcount
AS mutcount, COUNT(Analyzed_seqmuts.Diogenix) AS count
FROM Analyzed_seqmuts
GROUP BY Analyzed_seqmuts.diagnosis, Analyzed_seqmuts.mutcount
ORDER BY Analyzed_seqmuts.diagnosis, Analyzed_seqmuts.mutcount;

-- MSP repertoire
CREATE TABLE AnalyzeSequence2
AS SELECT SequencefilteredR1RMfinal.*
FROM AnalyzeRemove2
LEFT JOIN SequencefilteredR1RMfinal
ON (AnalyzeRemove2.batch = SequencefilteredR1RMfinal.batch AND
AnalyzeRemove2.sample = SequencefilteredR1RMfinal.sample);

CREATE TABLE Analyzemsp
AS SELECT AnalyzeSequence2.diagnosis AS diagnosis, AnalyzeSequence2.Diogenix
AS Diogenix, SUM(AnalyzeSequence2.RMcount) AS RMcount,
```

```
SUM(AnalyzeSequence2.AGS31B+AnalyzeSequence2.AGS40+AnalyzeSequence2.AG
S56+AnalyzeSequence2.AGS57+AnalyzeSequence2.AGS81) AS AGS5rm
FROM AnalyzeSequence2
GROUP BY AnalyzeSequence2.Diogenix
ORDER BY AnalyzeSequence2.diagnosis, AnalyzeSequence2.Diogenix;

CREATE TABLE AnalyzeOUTmsp
AS SELECT Analyzemsp.*, (Analyzemsp.AGS5rm*1.0 / Analyzemsp.RMcount*100 -
1.6*5) / 0.9 AS MSP
FROM Analyzemsp
ORDER BY Analyzemsp.diagnosis, Analyzemsp.Diogenix;
INSERT INTO AnalyzeOUTmsp
SELECT AnalyzeRemove3.diagnosis, AnalyzeRemove3.Diogenix, 0, 0, (0 - 1.6*5) / 0.9
FROM AnalyzeRemove3;
```

REFERENCES

Al-Lazikani B, Lesk AM, Chothia C (1997) Standard conformations for the canonical structures of immunoglobulins. Journal of molecular biology 273:927-948.

Alamyar E, Duroux P, Lefranc MP, Giudicelli V (2012) IMGT((R)) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. Methods in molecular biology 882:569-604.

Andersson M, Alvarez-Cermeno J, Bernardi G, Cogato I, Fredman P, Frederiksen J, Fredrikson S, Gallo P, Grimaldi LM, Gronning M, et al. (1994) Cerebrospinal fluid in the diagnosis of multiple sclerosis: a consensus report. J Neurol Neurosurg Psychiatry 57:897-902.

Angelucci F, Mirabella M, Frisullo G, Caggiula M, Tonali PA, Batocchi AP (2005) Serum levels of anti-myelin antibodies in relapsing-remitting multiple sclerosis patients during different phases of disease activity and immunomodulatory therapy. Disease markers 21:49-55.

Antel J, Bar-Or A (2006) Roles of immunoglobulins and B cells in multiple sclerosis: From pathogenesis to treatment. J Neuroimmunol 180:3-8.

Appel SH, Bornstein MB (1964) The Application of Tissue Culture to the Study of Experimental Allergic Encephalomyelitis. Ii. Serum Factors Responsible for Demyelination. J Exp Med 119:303-312.

Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, Nusbaum C, Rajewsky K, Koralov SB (2011) High-resolution description of antibody heavy-chain repertoires in humans. PloS one 6:e22365.

Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22:195-201.

Balzer S, Malde K, Lanzen A, Sharma A, Jonassen I (2010) Characteristics of 454 pyrosequencing data--enabling realistic simulation with flowsim. Bioinformatics 26:i420-425.

Baranzini SE, Jeong MC, Butunoi C, Murray RS, Bernard CC, Oksenberg JR (1999) B cell repertoire diversity and clonal expansion in multiple sclerosis brain lesions. J Immunol 163:5133-5144.

Bennett JL, Owens GP (2012) Cerebrospinal fluid proteomics: a new window for understanding human demyelinating disorders? Ann Neurol 71:587-588.

Bennett JL, Haubold K, Ritchie AM, Edwards SJ, Burgoon M, Shearer AJ, Gilden DH, Owens GP (2008) CSF IgG heavy-chain bias in patients at the time of a clinically isolated syndrome. J Neuroimmunol 199:126-132.

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Gallo Cassarino T, Bertoni M, Bordoli L, Schwede T (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic acids research 42:W252-258.

Bitsch A, Schuchardt J, Bunkowski S, Kuhlmann T, Bruck W (2000) Acute axonal injury in multiple sclerosis. Correlation with demyelination and inflammation. Brain 123 (Pt 6):1174-1183.

Bjartmar C, Kidd G, Mork S, Rudick R, Trapp BD (2000) Neurological disability correlates with spinal cord axonal loss and reduced N-acetyl aspartate in chronic multiple sclerosis patients. Ann Neurol 48:893-901.

Bjartmar C, Kinkel RP, Kidd G, Rudick RA, Trapp BD (2001) Axonal loss in normal-appearing white matter in a patient with acute MS. Neurology 57:1248-1252.

Bo L, Vedeler CA, Nyland HI, Trapp BD, Mork SJ (2003) Subpial demyelination in the cerebral cortex of multiple sclerosis patients. J Neuropathol Exp Neurol 62:723-732.

Bo L, Geurts JJ, van der Valk P, Polman C, Barkhof F (2007) Lack of correlation between cortical demyelination and white matter pathologic changes in multiple sclerosis. Arch Neurol 64:76-80.

Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV, Turchaninova MA, Lukyanov S, Lebedev YB, Chudakov DM (2012) Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. Eur J Immunol 42:3073-3083.

Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T (2008) Protein structure homology modeling using SWISS-MODEL workspace. Nat Protocols 4:1-13.

Boronat A, Sepulveda M, Llufriu S, Sabater L, Blanco Y, Gabilondo I, Sola N, Villoslada P, Dalmau J, Graus F, Saiz A (2012) Analysis of antibodies to surface epitopes of contactin-2 in multiple sclerosis. J Neuroimmunol 244:103-106.

Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, Nadeau KC, Egholm M, Miklos DB, Zehnder JL, Fire AZ (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Science translational medicine 1:12ra23.

Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, Nadeau KC, Egholm M, Miklos DB, Zehnder JL, Collins AM (2010) Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol 184:6986-6992.

Breij EC, Heijnen P, van der Goes A, Teunissen CE, Polman CH, Dijkstra CD (2006) Myelin flow cytometry assay detects enhanced levels of antibodies to human whole myelin in a subpopulation of multiple sclerosis patients. J Neuroimmunol 176:106-114.

Brex PA, Ciccarelli O, O'Riordan JI, Sailer M, Thompson AJ, Miller DH (2002) A longitudinal study of abnormalities on MRI and disability from multiple sclerosis. N Engl J Med 346:158-164.

Brezinschek HP, Brezinschek RI, Lipsky PE (1995) Analysis of the heavy chain repertoire of human peripheral B cells using single-cell polymerase chain reaction. J Immunol 155:190-202.

Brickshawana A, Hinson SR, Romero MF, Lucchinetti CF, Guo Y, Buttmann M, McKeon A, Pittock SJ, Chang MH, Chen AP, Kryzer TJ, Fryer JP, Jenkins SM, Cabre P, Lennon VA (2014) Investigation of the KIR4.1 potassium channel as a putative antigen in patients with multiple sclerosis: a comparative study. Lancet Neurol 13:795-806.

Brill L, Goldberg L, Karni A, Petrou P, Abramsky O, Ovadia H, Ben-Hur T, Karussis D, Vaknin-Dembinsky A (2015) Increased anti-KIR4.1 antibodies in multiple sclerosis: could it be a marker of disease relapse? Mult Scler 21:572-579.

Briney BS, Willis JR, McKinney BA, Crowe JE, Jr. (2012) High-throughput antibody sequencing reveals genetic evidence of global regulation of the naive and memory repertoires that extends across individuals. Genes and immunity 13:469-473.

Cameron EM, Spencer S, Lazarini J, Harp CT, Ward ES, Burgoon M, Owens GP, Racke MK, Bennett JL, Frohman EM, Monson NL (2009) Potential of a unique antibody gene signature to predict conversion to clinically definite multiple sclerosis. J Neuroimmunol 213:123-130.

Cepok S, von Geldern G, Grummel V, Hochgesand S, Celik H, Hartung H, Hemmer B (2006) Accumulation of class switched IgD-IgM- memory B cells in the cerebrospinal fluid during neuroinflammation. J Neuroimmunol 180:33-39.

Cepok S, Jacobsen M, Schock S, Omer B, Jaekel S, Boddeker I, Oertel WH, Sommer N, Hemmer B (2001) Patterns of cerebrospinal fluid pathology correlate with disease progression in multiple sclerosis. Brain 124:2169-2176.

Cepok S, Rosche B, Grummel V, Vogel F, Zhou D, Sayn J, Sommer N, Hartung HP, Hemmer B (2005) Short-lived plasma blasts are the main B cell effector subset during the course of multiple sclerosis. Brain 128:1667-1676.

Chard DT, Griffin CM, Rashid W, Davies GR, Altmann DR, Kapoor R, Barker GJ, Thompson AJ, Miller DH (2004) Progressive grey matter atrophy in clinically early relapsing-remitting multiple sclerosis. Mult Scler 10:387-391.

Chen D, Ireland SJ, Davis LS, Kong X, Stowe AM, Wang Y, White WI, Herbst R, Monson NL (2016) Autoreactive CD19+CD20- Plasma Cells Contribute to Disease Severity of Experimental Autoimmune Encephalomyelitis. J Immunol 196:1541-1549.

Chen D, Blazek M, Ireland S, Ortega S, Kong X, Meeuwissen A, Stowe A, Carter L, Wang Y, Herbst R, Monson NL (2014) Single dose of glycoengineered anti-CD19 antibody (MEDI551) disrupts experimental autoimmune encephalomyelitis by inhibiting pathogenic adaptive immune responses in the bone marrow and spinal cord while preserving peripheral regulatory mechanisms. J Immunol 193:4823-4832.

Chothia C, Lesk AM (1987) Canonical structures for the hypervariable regions of immunoglobulins. Journal of molecular biology 196:901-917.

Colombo M, Dono M, Gazzola P, Chiorazzi N, Mancardi G, Ferrarini M (2003) Maintenance of B lymphocyte-related clones in the cerebrospinal fluid of multiple sclerosis patients. Eur J Immunol 33:3433-3438.

Colombo M, Dono M, Gazzola P, Roncella S, Valetto A, Chiorazzi N, Mancardi GL, Ferrarini M (2000) Accumulation of clonally related B lymphocytes in the cerebrospinal fluid of multiple sclerosis patients. J Immunol 164:2782-2789.

Confavreux C, Vukusic S, Adeleine P (2003) Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. Brain 126:770-782.

Corcione A, Casazza S, Ferretti E, Giunti D, Zappia E, Pistorio A, Gambini C, Mancardi GL, Uccelli A, Pistoia V (2004) Recapitulation of B cell differentiation in the

central nervous system of patients with multiple sclerosis. Proceedings of the National Academy of Sciences of the United States of America 101:11064-11069.
Courtney AM, Treadaway K, Remington G, Frohman E (2009) Multiple sclerosis. Med Clin North Am 93:451-476, ix-x.
Cross AH, Stark JL, Lauber J, Ramsbottom MJ, Lyons JA (2006) Rituximab reduces B cells and T cells in cerebrospinal fluid of multiple sclerosis patients. J Neuroimmunol 180:63-70.
D'Alessandro R, Vignatelli L, Lugaresi A, Baldin E, Granella F, Tola MR, Malagu S, Motti L, Neri W, Galeotti M, Santangelo M, Fiorani L, Montanari E, Scandellari C, Benedetti MD, Leone M (2013) Risk of multiple sclerosis following clinically isolated syndrome: a 4-year prospective study. J Neurol 260:1583-1593.
Davidson CJ, Zeringer E, Champion KJ, Gauthier MP, Wang F, Boonyaratanakornkit J, Jones JR, Schreiber E (2012) Improving the limit of detection for Sanger sequencing: a comparison of methodologies for KRAS variant detection. BioTechniques 53:182-188.
de Graaf MT, Smitt PA, Luitwieler RL, van Velzen C, van den Broek PD, Kraan J, Gratama JW (2011) Central memory CD4+ T cells dominate the normal cerebrospinal fluid. Cytometry B Clin Cytom 80:43-50.
de Seze J, Lebrun C, Stojkovic T, Ferriby D, Chatel M, Vermersch P (2003) Is Devic's neuromyelitis optica a separate disease? A comparative study with multiple sclerosis. Multiple Sclerosis 9:521-525.
de Vries HE, Kooij G, Frenkel D, Georgopoulos S, Monsonego A, Janigro D (2012) Inflammatory events at blood-brain barrier in neuroinflammatory and neurodegenerative disorders: implications for clinical disease. Epilepsia 53 Suppl 6:45-52.
Derfuss T, Parikh K, Velhin S, Braun M, Mathey E, Krumbholz M, Kumpfel T, Moldenhauer A, Rader C, Sonderegger P, Pollmann W, Tiefenthaller C, Bauer J, Lassmann H, Wekerle H, Karagogeos D, Hohlfeld R, Linington C, Meinl E (2009) Contactin-2/TAG-1-directed autoimmunity is identified in multiple sclerosis patients and mediates gray matter pathology in animals. Proceedings of the National Academy of Sciences of the United States of America 106:8302-8307.
Dobson R, Ramagopalan S, Davis A, Giovannoni G (2013) Cerebrospinal fluid oligoclonal bands in multiple sclerosis and clinically isolated syndromes: a meta-analysis of prevalence, prognosis and effect of latitude. J Neurol Neurosurg Psychiatry.
Dorner T, Brezinschek HP, Foster SJ, Brezinschek RI, Farner NL, Lipsky PE (1998) Delineation of selective influences shaping the mutated expressed human Ig heavy chain repertoire. J Immunol 160:2831-2841.
Duddy M, Niino M, Adatia F, Hebert S, Freedman M, Atkins H, Kim HJ, Bar-Or A (2007) Distinct Effector Cytokine Profiles of Memory and Naive Human B Cell Subsets and Implication in Multiple Sclerosis. The Journal of Immunology 178:6092-6099.
Evdoshenko E, Maslyanskiy A, Lapin S, Zaslavsky L, Dobson R, Totolian A, Skoromets A, Bar-Or A (2013) Dynamics of B-Cell Populations in CSF and Blood in

Patients Treated with a Combination of Rituximab and Mitoxantrone. ISRN neurology 2013:748127.
Ferraro AJ, Drayson MT, Savage CO, MacLennan IC (2008) Levels of autoantibodies, unlike antibodies to all extrinsic antigen groups, fall following B cell depletion with Rituximab. Eur J Immunol 38:292-298.
Filina T, Feja KN, Tolan RW, Jr. (2013) An adolescent with pseudomigraine, transient headache, neurological deficits, and lymphocytic pleocytosis (HaNDL Syndrome): case report and review of the literature. Clinical pediatrics 52:496-502.
Fisher E, Rudick RA, Simon JH, Cutter G, Baier M, Lee JC, Miller D, Weinstock-Guttman B, Mass MK, Dougherty DS, Simonian NA (2002) Eight-year follow-up study of brain atrophy in patients with MS. Neurology 59:1412-1420.
Fisniku LK, Chard DT, Jackson JS, Anderson VM, Altmann DR, Miszkiel KA, Thompson AJ, Miller DH (2008a) Gray matter atrophy is related to long-term disability in multiple sclerosis. Ann Neurol 64:247-254.
Fisniku LK, Brex PA, Altmann DR, Miszkiel KA, Benton CE, Lanyon R, Thompson AJ, Miller DH (2008b) Disability and T2 MRI lesions: a 20-year follow-up of patients with relapse onset of multiple sclerosis. Brain 131:808-817.
Fraussen J, Vrolix K, Martinez-Martinez P, Losen M, De Baets MH, Stinissen P, Somers V (2009) B cell characterization and reactivity analysis in multiple sclerosis. Autoimmunity reviews 8:654-658.
Freedman MS, Thompson EJ, Deisenhammer F, Giovannoni G, Grimsley G, Keir G, Ohman S, Racke MK, Sharief M, Sindic CJ, Sellebjerg F, Tourtellotte WW (2005) Recommended standard of cerebrospinal fluid analysis in the diagnosis of multiple sclerosis: a consensus statement. Arch Neurol 62:865-870.
Frischer JM, Bramow S, Dal-Bianco A, Lucchinetti CF, Rauschka H, Schmidbauer M, Laursen H, Sorensen PS, Lassmann H (2009) The relation between inflammation and neurodegeneration in multiple sclerosis brains. Brain 132:1175-1189.
Frohman EM, Racke MK, Raine CS (2006a) Multiple sclerosis--the plaque and its pathogenesis. N Engl J Med 354:942-955.
Frohman EM, Havrdova E, Lublin F, Barkhof F, Achiron A, Sharief MK, Stuve O, Racke MK, Steinman L, Weiner H, Olek M, Zivadinov R, Corboy J, Raine C, Cutter G, Richert J, Filippi M (2006b) Most patients with multiple sclerosis or a clinically isolated demyelinating syndrome should be treated at the time of diagnosis. Arch Neurol 63:614-619.
Fuellgrabe MW, Herrmann D, Knecht H, Kuenzel S, Kneba M, Pott C, Bruggemann M (2015) High-Throughput, Amplicon-Based Sequencing of the CREBBP Gene as a Tool to Develop a Universal Platform-Independent Assay. PloS one 10:e0129195.
Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH (2015) Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. Proceedings of the National Academy of Sciences of the United States of America 112:E862-870.
Gajofatto A, Bongianni M, Zanusso G, Bianchi MR, Turatti M, Benedetti MD, Monaco S (2013) Clinical and biomarker assessment of demyelinating events suggesting multiple sclerosis. Acta Neurol Scand.

Galan M, Guivier E, Caraux G, Charbonnel N, Cosson JF (2010) A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. BMC genomics 11:296.
Gauld SB, Dal Porto JM, Cambier JC (2002) B cell antigen receptor signaling: roles in cell development and disease. Science 296:1641-1642.
Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. Nature biotechnology 32:158-168.
Geurts JJ, Bo L, Pouwels PJ, Castelijns JA, Polman CH, Barkhof F (2005) Cortical lesions in multiple sclerosis: combined postmortem MR imaging and histopathology. Ajnr 26:572-577.
Greenberg BM (2011) Treatment of acute transverse myelitis and its early complications. Continuum (Minneap Minn) 17:733-743.
Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH (2015) Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics 31:3356-3358.
Haas J, Bekeredjian-Ding I, Milkova M, Balint B, Schwarz A, Korporal M, Jarius S, Fritz B, Lorenz HM, Wildemann B (2011) B cells undergo unique compartmentalized redistribution in multiple sclerosis. J Autoimmun 37:289-299.
Hahn CD, Shroff MM, Blaser SI, Banwell BL (2004) MRI criteria for multiple sclerosis. Neurology 62:806-808.
Harp C, Lee J, Lambracht-Washington D, Cameron E, Olsen G, Frohman E, Racke M, Monson N (2007) Cerebrospinal fluid B cells from multiple sclerosis patients are subject to normal germinal center selection. J Neuroimmunol 183:189-199.
Harp CT, Ireland S, Davis LS, Remington G, Cassidy B, Cravens PD, Stuve O, Lovett-Racke AE, Eagar TN, Greenberg BM, Racke MK, Cowell LG, Karandikar NJ, Frohman EM, Monson NL (2010) Memory B cells from a subset of treatment-naïve relapsing-remitting multiple sclerosis patients elicit CD4+ T-cell proliferation and IFN-γ production in response to myelin basic protein and myelin oligodendrocyte glycoprotein. European Journal of Immunology 40:2942-2956.
Hauser SL, Waubant E, Arnold DL, Vollmer T, Antel J, Fox RJ, Bar-Or A, Panzara M, Sarkar N, Agarwal S, Langer-Gould A, Smith CH (2008) B-cell depletion with rituximab in relapsing-remitting multiple sclerosis. N Engl J Med 358:676-688.
Holman DW, Klein RS, Ransohoff RM (2011) The blood-brain barrier, chemokines and multiple sclerosis. Biochimica et biophysica acta 1812:220-230.
Ioannou Y, Lambrianides A, Cambridge G, Leandro MJ, Edwards JC, Isenberg DA (2008) B cell depletion therapy for patients with systemic lupus erythematosus results in a significant drop in anticardiolipin antibody titres. Ann Rheum Dis 67:425-426.
Ireland S, Monson N (2011) Potential impact of B cells on T cell function in multiple sclerosis. Mult Scler Int 2011:423971.
Ireland SJ, Blazek M, Harp CT, Greenberg B, Frohman EM, Davis LS, Monson NL (2012) Antibody-independent B cell effector functions in relapsing remitting multiple sclerosis: clues to increased inflammatory and reduced regulatory B cell capacity. Autoimmunity 45:400-414.

Ireland SJ, Guzman AA, O'Brien DE, Hughes S, Greenberg B, Flores A, Graves D, Remington G, Frohman EM, Davis LS, Monson NL (2014) The effect of glatiramer acetate therapy on functional properties of B cells from patients with relapsing-remitting multiple sclerosis. JAMA neurology 71:1421-1428.
Izquierdo G, Angulo S, Garcia-Moreno JM, Gamero MA, Navarro G, Gata JM, Ruiz-Pena JL, Paramo MD (2002) Intrathecal IgG synthesis: marker of progression in multiple sclerosis patients. Acta Neurol Scand 105:158-163.
Jackson KJ, Kidd MJ, Wang Y, Collins AM (2013) The Shape of the Lymphocyte Receptor Repertoire: Lessons from the B Cell Receptor. Front Immunol 4:263.
Jacobs LD, Kaba SE, Miller CM, Priore RL, Brownscheidle CM (1997) Correlation of clinical, magnetic resonance imaging, and cerebrospinal fluid findings in optic neuritis. Ann Neurol 41:392-398.
Jiang Y, Nie K, Redmond D, Melnick AM, Tam W, Elemento O (2015) VDJ-Seq: Deep Sequencing Analysis of Rearranged Immunoglobulin Heavy Chain Gene to Reveal Clonal Evolution Patterns of B Cell Lymphoma. Journal of visualized experiments : JoVE:e53215.
Johnson KP, Nelson BJ (1977) Multiple sclerosis: diagnostic usefulness of cerebrospinal fluid. Ann Neurol 2:425-431.
Kabat EA, Moore DH, Landow H (1942) An Electrophoretic Study of the Protein Components in Cerebrospinal Fluid and Their Relationship to the Serum Proteins. The Journal of clinical investigation 21:571-577.
Kabat EA, Glusman M, Knaub V (1948) Quantitative Estimation of the Albumin and Gamma Globulin in Normal and Pathologic Cerebrospinal Fluid by Immunochemical Methods. Am J Med 4:653-662.
Kabat EA, Freedman DA, et al. (1950) A study of the crystalline albumin, gamma globulin and total protein in the cerebrospinal fluid of 100 cases of multiple sclerosis and in other diseases. Am J Med Sci 219:55-64.
Kabat EA, Te Wu T, Perry HM, Gottesman KS, Foeller C (1992) Sequences of proteins of immunological interest: DIANE publishing.
Kappos L, Radue EW, O'Connor P, Polman C, Hohlfeld R, Calabresi P, Selmaj K, Agoropoulou C, Leyk M, Zhang-Auberson L, Burtin P (2010) A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. N Engl J Med 362:387-401.
Kappos L, Li D, Calabresi PA, O'Connor P, Bar-Or A, Barkhof F, Yin M, Leppert D, Glanzman R, Tinbergen J, Hauser SL (2011) Ocrelizumab in relapsing-remitting multiple sclerosis: a phase 2, randomised, placebo-controlled, multicentre trial. Lancet 378:1779-1787.
Karni A, Bakimer-Kleiner R, Abramsky O, Ben-Nun A (1999) Elevated levels of antibody to myelin oligodendrocyte glycoprotein is not specific for patients with multiple sclerosis. Arch Neurol 56:311-315.
Kawamura N, Yamasaki R, Yonekawa T, Matsushita T, Kusunoki S, Nagayama S, Fukuda Y, Ogata H, Matsuse D, Murai H, Kira J (2013) Anti-neurofascin antibody in patients with combined central and peripheral demyelination. Neurology 81:714-722.
Keegan M, Konig F, McClelland R, Bruck W, Morales Y, Bitsch A, Panitch H, Lassmann H, Weinshenker B, Rodriguez M, Parisi J, Lucchinetti CF (2005)

Relation between humoral pathological changes in multiple sclerosis and response to therapeutic plasma exchange. Lancet 366:579-582.
Khan TA, Friedensohn S, Gorter de Vries AR, Straszewski J, Ruscheweyh HJ, Reddy ST (2016) Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. Science advances 2:e1501371.
Kidd D, Barkhof F, McConnell R, Algra PR, Allen IV, Revesz T (1999) Cortical lesions in multiple sclerosis. Brain 122 (Pt 1):17-26.
Kim SH, Kim W, Li XF, Jung IJ, Kim HJ (2011) Repeated treatment with rituximab based on the assessment of peripheral circulating memory B cells in patients with relapsing neuromyelitis optica over 2 years. Arch Neurol 68:1412-1420.
Kircher M, Kelso J (2010) High-throughput DNA sequencing--concepts and limitations. BioEssays : news and reviews in molecular, cellular and developmental biology 32:524-536.
Kirk J, Plumb J, Mirakhur M, McQuaid S (2003) Tight junctional abnormality in multiple sclerosis white matter affects all calibres of vessel and is associated with blood-brain barrier leakage and active demyelination. J Pathol 201:319-327.
Klawiter EC, Piccio L, Lyons JA, Mikesell R, O'Connor KC, Cross AH (2010) Elevated intrathecal myelin oligodendrocyte glycoprotein antibodies in multiple sclerosis. Arch Neurol 67:1102-1108.
Kleine TO, Benes L (2006) Immune surveillance of the human central nervous system (CNS): different migration pathways of immune cells through the blood-brain barrier and blood-cerebrospinal fluid barrier in healthy persons. Cytometry A 69:147-151.
Kowarik MC, Pellkofer HL, Cepok S, Korn T, Kumpfel T, Buck D, Hohlfeld R, Berthele A, Hemmer B (2011) Differential effects of fingolimod (FTY720) on immune cells in the CSF and blood of patients with MS. Neurology 76:1214-1221.
Kroksveen AC, Opsahl JA, Guldbrandsen A, Myhr K, Oveland E, Torkildsen O, Berven FS (2014) Cerebrospinal fluid proteomics in multiple sclerosis. Biochimica et biophysica acta.
Krumbholz M, Derfuss T, Hohlfeld R, Meinl E (2012) B cells and antibodies in multiple sclerosis pathogenesis and therapy. Nat Rev Neurol 8:613-623.
Kuhle J, Pohl C, Mehling M, Edan G, Freedman MS, Hartung HP, Polman CH, Miller DH, Montalban X, Barkhof F, Bauer L, Dahms S, Lindberg R, Kappos L, Sandbrink R (2007) Lack of association between antimyelin antibodies and progression to multiple sclerosis. N Engl J Med 356:371-378.
Lampasona V, Franciotta D, Furlan R, Zanaboni S, Fazio R, Bonifacio E, Comi G, Martino G (2004) Similar low frequency of anti-MOG IgG and IgM in MS patients and healthy subjects. Neurology 62:2092-2094.
Lassmann H, Bruck W, Lucchinetti CF (2007) The immunopathology of multiple sclerosis: an overview. Brain Pathol 17:210-218.
Lazarus MN, Turner-Stokes T, Chavele KM, Isenberg DA, Ehrenstein MR (2012) B-cell numbers and phenotype at clinical relapse following rituximab therapy differ in SLE patients according to anti-dsDNA antibody levels. Rheumatology (Oxford) 51:1208-1215.

Leech S, Kirk J, Plumb J, McQuaid S (2007) Persistent endothelial abnormalities and blood-brain barrier leak in primary and secondary progressive multiple sclerosis. Neuropathol Appl Neurobiol 33:86-98.
Lefranc MP (2003) IMGT, the international ImMunoGeneTics database(R). Nucleic acids research 31:307-310.
Lennon V, Wingerchuk D, Kryzer T, Pittock S, Lucchinetti C, Fujihara K, Nakashima I, Weinshenker B (2004) A serum autoantibody marker of neuromyelitis optica: distinction from multiple sclerosis. The Lancet 364:2106-2112.
Lennon VA, Kryzer TJ, Pittock SJ, Verkman AS, Hinson SR (2005) IgG marker of optic-spinal multiple sclerosis binds to the aquaporin-4 water channel. The Journal of Experimental Medicine 202:473-477.
Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady.
Ligocki AJ (2014) Implications of dysregulated antibody hypermutation patterns in multiple sclerosis patients (Doctoral dissertation).
Ligocki AJ, Rounds WH, Cameron EM, Harp CT, Frohman EM, Courtney AM, Vernino S, Cowell LG, Greenberg B, Monson NL (2013) Expansion of CD27high plasmablasts in transverse myelitis patients that utilize VH4 and JH6 genes and undergo extensive somatic hypermutation. Genes and immunity 14:291-301.
Ligocki AJ, Lovato L, Xiang D, Guidry P, Scheuermann RH, Willis SN, Almendinger S, Racke MK, Frohman EM, Hafler DA, O'Connor KC, Monson NL (2010) A unique antibody gene signature is prevalent in the central nervous system of patients with multiple sclerosis. J Neuroimmunol 226:192-193.
Ligocki AJ, Rivas JR, Rounds WH, Guzman AA, Li M, Spadaro M, Lahey L, Chen D, Henson PM, Graves D, Greenberg BM, Frohman EM, Ward ES, Robinson W, Meinl E, White CL, 3rd, Stowe AM, Monson NL (2015) A Distinct Class of Antibodies May Be an Indicator of Gray Matter Autoimmunity in Early and Established Relapsing Remitting Multiple Sclerosis Patients. ASN neuro 7.
Lindner M, Ng JK, Hochmeister S, Meinl E, Linington C (2013) Neurofascin 186 specific autoantibodies induce axonal injury and exacerbate disease severity in experimental autoimmune encephalomyelitis. Exp Neurol.
Linington C, Bradl M, Lassmann H, Brunner C, Vass K (1988) Augmentation of demyelination in rat acute allergic encephalomyelitis by circulating mouse monoclonal antibodies directed against a myelin/oligodendrocyte glycoprotein. Am J Pathol 130:443-454.
Link H, Muller R (1971) Immunoglobulins in multiple sclerosis and infections of the nervous system. Arch Neurol 25:326-344.
Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, Buno I, Armstrong R, Fire AZ, Weinberg KI, Mindrinos M, Zehnder JL, Boyd SD, Xiao W, Davis RW, Miklos DB (2011) High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. Proceedings of the National Academy of Sciences of the United States of America 108:21194-21199.
Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nature biotechnology 30:434-439.

Lublin FD, Reingold SC (1996) Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Neurology 46:907-911.

Lucchinetti C, Bruck W, Parisi J, Scheithauer B, Rodriguez M, Lassmann H (2000) Heterogeneity of multiple sclerosis lesions: implications for the pathogenesis of demyelination. Ann Neurol 47:707-717.

Lucchinetti CF, Mandler RN, McGavern D, Bruck W, Gleich G, Ransohoff RM, Trebst C, Weinshenker B, Wingerchuk D, Parisi JE, Lassmann H (2002) A role for humoral mechanisms in the pathogenesis of Devic's neuromyelitis optica. Brain 125:1450-1461.

Lucchinetti CF, Popescu BF, Bunyan RF, Moll NM, Roemer SF, Lassmann H, Bruck W, Parisi JE, Scheithauer BW, Giannini C, Weigand SD, Mandrekar J, Ransohoff RM (2011) Inflammatory cortical demyelination in early multiple sclerosis. N Engl J Med 365:2188-2197.

Luxton RW, McLean BN, Thompson EJ (1990) Isoelectric focusing versus quantitative measurements in the detection of intrathecal local synthesis of IgG. Clin Chim Acta 187:297-308.

Magana SM, Keegan BM, Weinshenker BG, Erickson BJ, Pittock SJ, Lennon VA, Rodriguez M, Thomsen K, Weigand S, Mandrekar J, Linbo L, Lucchinetti CF (2011) Beneficial plasma exchange response in central nervous system inflammatory demyelination. Arch Neurol 68:870-878.

Magliozzi R, Howell O, Vora A, Serafini B, Nicholas R, Puopolo M, Reynolds R, Aloisi F (2007) Meningeal B-cell follicles in secondary progressive multiple sclerosis associate with early onset of disease and severe cortical pathology. Brain 130:1089-1104.

Mardis ER (2011) A decade's perspective on DNA sequencing technology. Nature 470:198-203.

Martin Mdel P, Cravens PD, Winger R, Kieseier BC, Cepok S, Eagar TN, Zamvil SS, Weber MS, Frohman EM, Kleinschmidt-Demasters BK, Montine TJ, Hemmer B, Marra CM, Stuve O (2009) Depletion of B lymphocytes from cerebral perivascular spaces by rituximab. Arch Neurol 66:1016-1020.

Mathey EK, Derfuss T, Storch MK, Williams KR, Hales K, Woolley DR, Al-Hayani A, Davies SN, Rasband MN, Olsson T, Moldenhauer A, Velhin S, Hohlfeld R, Meinl E, Linington C (2007) Neurofascin as a novel target for autoantibody-mediated axonal injury. J Exp Med 204:2363-2372.

McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD, McFarland HF, Paty DW, Polman CH, Reingold SC, Sandberg-Wollheim M, Sibley W, Thompson A, van den Noort S, Weinshenker BY, Wolinsky JS (2001) Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 50:121-127.

Meffre E, Casellas R, Nussenzweig MC (2000) Antibody regulation of B cell development. Nature immunology 1:379-385.

Meffre E, Chiorazzi M, Nussenzweig MC (2001) Circulating human B cells that express surrogate light chains display a unique antibody repertoire. J Immunol 167:2151-2156.
Miller D, Barkhof F, Montalban X, Thompson A, Filippi M (2005) Clinically isolated syndromes suggestive of multiple sclerosis, part I: natural history, pathogenesis, diagnosis, and prognosis. Lancet Neurol 4:281-288.
Miller DH, Rudge P, Johnson G, Kendall BE, Macmanus DG, Moseley IF, Barnes D, McDonald WI (1988) Serial gadolinium enhanced magnetic resonance imaging in multiple sclerosis. Brain 111 (Pt 4):927-939.
Milo R, Miller A (2014) Revised diagnostic criteria of multiple sclerosis. Autoimmunity reviews.
Minagar A, Alexander JS (2003) Blood-brain barrier disruption in multiple sclerosis. Mult Scler 9:540-549.
Molyneux PD, Filippi M, Barkhof F, Gasperini C, Yousry TA, Truyen L, Lai HM, Rocca MA, Moseley IF, Miller DH (1998) Correlations between monthly enhanced MRI lesion rate and changes in T2 lesion volume in multiple sclerosis. Ann Neurol 43:332-339.
Monson NL, Cravens PD, Frohman EM, Hawker K, Racke MK (2005a) Effect of Rituximab on the peripheral blood and cerebrospinal fluid b cells in patients with primary progressive multiple sclerosis. Archives of Neurology 62:258-264.
Monson NL, Brezinschek HP, Brezinschek RI, Mobley A, Vaughan GK, Frohman EM, Racke MK, Lipsky PE (2005b) Receptor revision and atypical mutational characteristics in clonally expanded B cells from the cerebrospinal fluid of recently diagnosed multiple sclerosis patients. J Neuroimmunol 158:170-181.
Moore F, Okuda DT (2009) Incidental MRI anomalies suggestive of multiple sclerosis: the radiologically isolated syndrome. Neurology 73:1714.
Moriarty DM, Blackshaw AJ, Talbot PR, Griffiths HL, Snowden JS, Hillier VF, Capener S, Laitt RD, Jackson A (1999) Memory dysfunction in multiple sclerosis corresponds to juxtacortical lesion load on fast fluid-attenuated inversion-recovery MR images. Ajnr 20:1956-1962.
Nerrant E, Salsac C, Charif M, Ayrignac X, Carra-Dalliere C, Castelnovo G, Goulabchand R, Tisseyre J, Raoul C, Eliaou JF, Labauge P, Vincent T (2014) Lack of confirmation of anti-inward rectifying potassium channel 4.1 antibodies as reliable markers of multiple sclerosis. Mult Scler 20:1699-1703.
Noseworthy JH, Lucchinetti C, Rodriguez M, Weinshenker BG (2000) Multiple sclerosis. N Engl J Med 343:938-952.
O'Connor KC, Chitnis T, Griffin DE, Piyasirisilp S, Bar-Or A, Khoury S, Wucherpfennig KW, Hafler DA (2003) Myelin basic protein-reactive autoantibodies in the serum and cerebrospinal fluid of multiple sclerosis patients are characterized by low-affinity interactions. J Neuroimmunol 136:140-148.
Obermeier B, Lovato L, Mentele R, Bruck W, Forne I, Imhof A, Lottspeich F, Turk KW, Willis SN, Wekerle H, Hohlfeld R, Hafler DA, O'Connor KC, Dornmair K (2011) Related B cell clones that populate the CSF and CNS of patients with multiple sclerosis produce CSF immunoglobulin. J Neuroimmunol 233:245-248.
Owens GP, Bennett JL, Gilden DH, Burgoon MP (2006) The B cell response in multiple sclerosis. Neurological research 28:236-244.

Owens GP, Burgoon MP, Anthony J, Kleinschmidt-DeMasters BK, Gilden DH (2001) The immunoglobulin G heavy chain repertoire in multiple sclerosis plaques is distinct from the heavy chain repertoire in peripheral blood lymphocytes. Clin Immunol 98:258-263.

Owens GP, Kraus H, Burgoon MP, Smith-Jensen T, Devlin ME, Gilden DH (1998) Restricted use of VH4 germline segments in an acute multiple sclerosis brain. Ann Neurol 43:236-243.

Owens GP, Ritchie AM, Burgoon MP, Williamson RA, Corboy JR, Gilden DH (2003) Single-cell repertoire analysis demonstrates that clonal expansion is a prominent feature of the B cell response in multiple sclerosis cerebrospinal fluid. J Immunol 171:2725-2733.

Owens GP, Winges KM, Ritchie AM, Edwards S, Burgoon MP, Lehnhoff L, Nielsen K, Corboy J, Gilden DH, Bennett JL (2007) VH4 gene segments dominate the intrathecal humoral immune response in multiple sclerosis. J Immunol 179:6343-6351.

Palanichamy A, Apeltsin L, Kuo TC, Sirota M, Wang S, Pitts SJ, Sundar PD, Telman D, Zhao LZ, Derstine M, Abounasr A, Hauser SL, von Budingen HC (2014) Immunoglobulin class-switched B cells form an active immune axis between CNS and periphery in multiple sclerosis. Science translational medicine 6:248ra106.

Petzold A (2013) Intrathecal oligoclonal IgG synthesis in multiple sclerosis. J Neuroimmunol 262:1-10.

Piccio L, Naismith RT, Trinkaus K, Klein RS, Parks BJ, Lyons JA, Cross AH (2010) Changes in B- and T-lymphocyte and chemokine levels with rituximab treatment in multiple sclerosis. Arch Neurol 67:707-714.

Pieterse V, Black PE (eds. 22 August 2013; Available from: http://www.nist.gov/dads/HTML/Levenshtein.html) Algorithms and Theory of Computation Handbook, CRC Press LLC, 1999, "Levenshtein distance". Dictionary of Algorithms and Data Structures [online].

Polman CH, Reingold SC, Banwell B, Clanet M, Cohen JA, Filippi M, Fujihara K, Havrdova E, Hutchinson M, Kappos L, Lublin FD, Montalban X, O'Connor P, Sandberg-Wollheim M, Thompson AJ, Waubant E, Weinshenker B, Wolinsky JS (2011) Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann Neurol 69:292-302.

Prabakaran P, Streaker E, Chen W, Dimitrov DS (2011) 454 antibody sequencing - error characterization and correction. BMC research notes 4:404.

Qin Y, Duquette P, Zhang Y, Talbot P, Poole R, Antel J (1998) Clonal expansion and somatic hypermutation of V(H) genes of B cells from cerebrospinal fluid in multiple sclerosis. The Journal of clinical investigation 102:1045-1050.

Qin Y, Duquette P, Zhang Y, Olek M, Da RR, Richardson J, Antel JP, Talbot P, Cashman NR, Tourtellotte WW, Wekerle H, Van Den Noort S (2003) Intrathecal B-cell clonal expansion, an early sign of humoral immunity, in the cerebrospinal fluid of patients with clinically isolated syndrome suggestive of multiple sclerosis. Lab Invest 83:1081-1088.

Quail MA, Smith M, Jackson D, Leonard S, Skelly T, Swerdlow HP, Gu Y, Ellis P (2014) SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. BMC genomics 15:110.

Quarles RH (2005) Comparison of CNS and PNS myelin proteins in the pathology of myelin disorders. J Neurol Sci 228:187-189.
Reske D, Petereit HF, Heiss WD (2005) Difficulties in the differentiation of chronic inflammatory diseases of the central nervous system--value of cerebrospinal fluid analysis and immunological abnormalities in the diagnosis. Acta Neurol Scand 112:207-213.
Ritchie AM, Gilden DH, Williamson RA, Burgoon MP, Yu X, Helm K, Corboy JR, Owens GP (2004) Comparative analysis of the CD19+ and CD138+ cell antibody repertoires in the cerebrospinal fluid of patients with multiple sclerosis. J Immunol 173:649-656.
Rocca MA, Agosta F, Sormani MP, Fernando K, Tintore M, Korteweg T, Tortorella P, Miller DH, Thompson A, Rovira A, Montalban X, Polman C, Barkhof F, Filippi M (2008) A three-year, multi-parametric MRI study in patients at presentation with CIS. J Neurol 255:683-691.
Rommer PS, Patejdl R, Winkelmann A, Benecke R, Zettl UK (2011) Rituximab for secondary progressive multiple sclerosis: a case series. CNS Drugs 25:607-613.
Rounds WH, Ligocki AJ, Levin MK, Greenberg BM, Bigwood DW, Eastman EM, Cowell LG, Monson NL (2014) The antibody genetics of multiple sclerosis: comparing next-generation sequencing to sanger sequencing. Frontiers in neurology 5:166.
Rounds WH, Salinas EA, Wilks TB, 2nd, Levin MK, Ligocki AJ, Ionete C, Pardo CA, Vernino S, Greenberg BM, Bigwood DW, Eastman EM, Cowell LG, Monson NL (2015) MSPrecise: A molecular diagnostic test for multiple sclerosis using next generation sequencing. Gene 572:191-197.
Sadaba MC, Tzartos J, Paino C, Garcia-Villanueva M, Alvarez-Cermeno JC, Villar LM, Esiri MM (2012) Axonal and oligodendrocyte-localized IgM and IgG deposits in MS lesions. J Neuroimmunol 247:86-94.
Scalfari A, Neuhaus A, Degenhardt A, Rice GP, Muraro PA, Daumer M, Ebers GC (2010) The natural history of multiple sclerosis: a geographically based study 10: relapses and long-term disability. Brain 133:1914-1929.
Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic acids research 43:e37.
Schluesener HJ, Sobel RA, Linington C, Weiner HL (1987) A monoclonal antibody against a myelin oligodendrocyte glycoprotein induces relapses and demyelination in central nervous system autoimmune disease. J Immunol 139:4016-4021.
Schrodinger, LLC (2015) The PyMOL Molecular Graphics System, Version 1.8. In.
Seil FJ, Falk GA, Kies MW, Alvord EC, Jr. (1968) The in vitro demyelinating activity of sera from guinea pigs sensitized with whole CNS and with purified encephalitogen. Exp Neurol 22:545-555.
Seitz V, Schaper S, Droge A, Lenze D, Hummel M, Hennig S (2015) A new method to prevent carry-over contaminations in two-step PCR NGS library preparations. Nucleic acids research 43:e135.
Sellebjerg F, Jensen CV, Christiansen M (2000) Intrathecal IgG synthesis and autoantibody-secreting cells in multiple sclerosis. J Neuroimmunol 108:207-215.

Sellebjerg F, Jaliashvili I, Christiansen M, Garred P (1998) Intrathecal activation of the complement system and disability in multiple sclerosis. J Neurol Sci 157:168-174.
Serafini B, Rosicarelli B, Magliozzi R, Stigliano E, Aloisi F (2004) Detection of ectopic B-cell follicles with germinal centers in the meninges of patients with secondary progressive multiple sclerosis. Brain Pathol 14:164-174.
Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature biotechnology 26:1135-1145.
Smith-Jensen T, Burgoon MP, Anthony J, Kraus H, Gilden DH, Owens GP (2000) Comparison of immunoglobulin G heavy-chain sequences in MS and SSPE brains reveals an antigen-driven response. Neurology 54:1227-1232.
Sorensen PS, Lisby S, Grove R, Derosier F, Shackelford S, Havrdova E, Drulovic J, Filippi M (2014) Safety and efficacy of ofatumumab in relapsing-remitting multiple sclerosis: a phase 2 study. Neurology 82:573-581.
Sormani MP, Arnold DL, De Stefano N (2013) Treatment effect on brain atrophy correlates with treatment effect on disability in multiple sclerosis. Ann Neurol.
Srivastava R, Aslam M, Kalluri SR, Schirmer L, Buck D, Tackenberg B, Rothhammer V, Chan A, Gold R, Berthele A, Bennett JL, Korn T, Hemmer B (2012) Potassium channel KIR4.1 as an immune target in multiple sclerosis. N Engl J Med 367:115-123.
Stuve O, Bennett JL (2007) Pharmacological properties, toxicology and scientific rationale for the use of natalizumab (Tysabri) in inflammatory diseases. CNS Drug Rev 13:79-95.
Stuve O, Marra CM, Jerome KR, Cook L, Cravens PD, Cepok S, Frohman EM, Phillips JT, Arendt G, Hemmer B, Monson NL, Racke MK (2006) Immune surveillance in multiple sclerosis patients treated with natalizumab. Ann Neurol 59:743-747.
Svenningsson A, Andersen O, Edsbagge M, Stemme S (1995) Lymphocyte phenotype and subset distribution in normal cerebrospinal fluid. J Neuroimmunol 63:39-46.
Tewarie P, Teunissen CE, Dijkstra CD, Heijnen DA, Vogt M, Balk L, Vrenken H, Polman CH, Killestein J (2012) Cerebrospinal fluid anti-whole myelin antibodies are not correlated to magnetic resonance imaging activity in multiple sclerosis. J Neuroimmunol 251:103-106.
Thrower BW (2007) Clinically isolated syndromes: predicting and delaying multiple sclerosis. Neurology 68:S12-15.
Tiller T, Meffre E, Yurasov S, Tsuiji M, Nussenzweig MC, Wardemann H (2008) Efficient generation of monoclonal antibodies from single human B cells by single cell RT-PCR and expression vector cloning. J Immunol Methods 329:112-124.
Tintore M, Rovira A, Rio J, Tur C, Pelayo R, Nos C, Tellez N, Perkal H, Comabella M, Sastre-Garriga J, Montalban X (2008) Do oligoclonal bands add information to MRI in first attacks of multiple sclerosis? Neurology 70:1079-1083.
Trapp BD, Peterson J, Ransohoff RM, Rudick R, Mork S, Bo L (1998) Axonal transection in the lesions of multiple sclerosis. N Engl J Med 338:278-285.
Uzawa A, Mori M, Hayakawa S, Masuda S, Kuwabara S (2010) Different responses to interferon beta-1b treatment in patients with neuromyelitis optica and multiple sclerosis. European Journal of Neurology 17:672-676.

Valsasina P, Benedetti B, Rovaris M, Sormani MP, Comi G, Filippi M (2005) Evidence for progressive gray matter loss in patients with relapsing-remitting MS. Neurology 65:1126-1128.
van Dongen JJ et al. (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia 17:2257-2317.
Villar LM, Garcia-Sanchez MI, Costa-Frossard L, Espino M, Roldan E, Paramo D, Lucas M, Izquierdo G, Alvarez-Cermeno JC (2012) Immunological markers of optimal response to natalizumab in multiple sclerosis. Arch Neurol 69:191-197.
von Budingen HC, Harrer MD, Kuenzle S, Meier M, Goebels N (2008) Clonally expanded plasma cells in the cerebrospinal fluid of MS patients produce myelin-specific antibodies. Eur J Immunol 38:2014-2023.
von Budingen HC, Gulati M, Kuenzle S, Fischer K, Rupprecht TA, Goebels N (2010) Clonally expanded plasma cells in the cerebrospinal fluid of patients with central nervous system autoimmune demyelination produce "oligoclonal bands". J Neuroimmunol 218:134-139.
von Budingen HC, Kuo TC, Sirota M, van Belle CJ, Apeltsin L, Glanville J, Cree BA, Gourraud PA, Schwartzburg A, Huerta G, Telman D, Sundar PD, Casey T, Cox DR, Hauser SL (2012) B cell exchange across the blood-brain barrier in multiple sclerosis. The Journal of clinical investigation 122:4533-4543.
Wang HY, Matsui M, Saida T (2002) Immunological disturbances in the central nervous system linked to MRI findings in multiple sclerosis. J Neuroimmunol 125:149-154.
Wang Y, Jackson KJ, Sewell WA, Collins AM (2008) Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. Immunology and cell biology 86:111-115.
Wang Y, Jackson KJ, Gaeta B, Pomat W, Siba P, Sewell WA, Collins AM (2011) Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. Immunogenetics 63:259-265.
Watson CT, Breden F (2012) The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes and immunity 13:363-373.
Wingerchuk DM, Hogancamp WF, O'Brien PC, Weinshenker BG (1999) The clinical course of neuromyelitis optica (Devic's syndrome). Neurology 53:1107-1114.
Wingerchuk DM, Lennon VA, Pittock SJ, Lucchinetti CF, Weinshenker BG (2006) Revised diagnostic criteria for neuromyelitis optica. Neurology 66:1485-1489.
Winges KM, Gilden DH, Bennett JL, Yu X, Ritchie AM, Owens GP (2007) Analysis of multiple sclerosis cerebrospinal fluid reveals a continuum of clonally related antibody-secreting cells that are predominantly plasma blasts. J Neuroimmunol 192:226-234.
Yaari G, Benichou JI, Vander Heiden JA, Kleinstein SH, Louzoun Y (2015) The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. Philosophical transactions of the Royal Society of London Series B, Biological sciences 370.

Ye J, Ma N, Madden TL, Ostell JM (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic acids research 41:W34-40.