

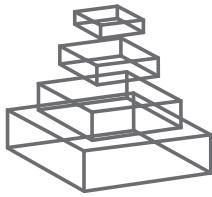
frontiers RESEARCH TOPICS

THE EVOLUTION AND
DEVELOPMENT OF THE
ANTIBODY REPERTOIRE

Topic Editor
Harry W. Schroeder Jr.



frontiers in
IMMUNOLOGY



FRONTIERS COPYRIGHT STATEMENT

© Copyright 2007-2015
Frontiers Media SA.
All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714

ISBN 978-2-88919-549-7

DOI 10.3389/978-2-88919-549-7

ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.

All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

DEDICATION TO QUALITY

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

THE EVOLUTION AND DEVELOPMENT OF THE ANTIBODY REPERTOIRE

Topic Editor:

Harry W. Schroeder Jr., University of Alabama at Birmingham, USA

Although at first glance mechanisms used to create the variable domains of immunoglobulin appear to be designed to generate diversity at random, closer inspection reveals striking evolutionary constraints on the sequence and structure of these antigen receptors, suggesting that natural selection is operating to create a repertoire that anticipates or is biased towards recognition of specific antigenic properties. This Research Topics issue will be devoted to an examination of the evolution of antigen receptor sequence at the germline level, an evaluation of the repertoire in B cells from fish, pigs and human, an introduction into bioinformatics approaches to the evaluation and analysis of the repertoire as ascertained by high throughput sequencing, and a discussion of how study of the normal repertoire informs the construction or selection of in vitro antibodies for applied purposes.

Table of Contents

- 04 *The evolution and development of the antibody repertoire***
Harry W. Schroeder Jr.
- 06 *The astonishing diversity of Ig classes and B cell repertoires in teleost fish***
Simon Fillatreau, Adrien Six, Susanna Magadan, Rosario Castro, J. Oriol Sunyer and Pierre Boudinot
- 20 *The porcine antibody repertoire: variations on the textbook theme***
John E. Butler and Nancy Wertz
- 34 *Fundamental roles of the innate-like repertoire of natural antibodies in immune homeostasis***
Jaya Vas, Caroline Grönwall and Gregg J. Silverman
- 42 *Differences in the composition of the human antibody repertoire by B cell subsets in the blood***
Eva Szymanska Mrocze, Gregory C. Ippolito, Tobias Rogosch, Kam Hon Hoi, Tracy A. Hwangpo, Marsha G. Brand, Yingxin Zhuang, Cun Ren Liu, David A. Schneider, Michael Zemlin, Elizabeth E. Brown, George Georgiou and Harry W. Schroeder Jr.
- 56 *Secondary mechanisms of diversification in the human antibody repertoire***
Bryan S. Briney and James E. Crowe Jr.
- 63 *Age-related changes in human peripheral blood IGH repertoire following vaccination***
Yu-Chang Bryan Wu, David Kipling and Deborah K. Dunn-Walters
- 75 *Natural and man-made V-gene repertoires for antibody discovery***
William J. J. Finlay and Juan C. Almagro
- 93 *Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing***
Miri Michaeli, Hila Noga, Hilla Tabibian-Keissar, Iris Barshack and Ramit Mehr
- 109 *Immunoglobulin Analysis Tool: a novel tool for the analysis of human and mouse heavy and light chain transcripts***
Tobias Rogosch, Sebastian Kerzel, Kam Hon Hoi, Zhixin Zhang, Rolf F. Maier, Gregory C. Ippolito and Michael Zemlin



The evolution and development of the antibody repertoire

Harry W. Schroeder Jr.*

Department of Medicine, Division of Clinical Immunology and Rheumatology, University of Alabama at Birmingham, Birmingham, AL, USA

*Correspondence: hwsj@uab.edu

Edited and reviewed by:

Thomas L. Rothstein, The Feinstein Institute for Medical Research, USA

Keywords: immunoglobulin, antibody repertoire, comparative immunology, developmental immunology, high-throughput sequencing

Approximately 500 million years ago (1), vertebrates developed the ability to generate a highly diverse repertoire of immunoglobulins (Igs). These highly versatile proteins serve as both effector molecules and as receptors for antigen ligands. As soluble effectors, Igs can activate and fix complement and they can bind Fc receptors on the surfaces of granulocytes, monocytes, platelets, and other components of the immune response. V(D)J gene segment rearrangement and somatic hypermutation (SHM) create a population of diverse ligand binding sites that allow recognition of an almost unlimited array of self and non-self antigens. Above and beyond the time-honored practice of vaccination, the power of Igs as biotherapeutic agents is changing the face of medicine. In this research topic, we collected several exciting articles that highlight the diversity and similarity of antibody repertoires. We also highlight new bioinformatics approaches for the analysis of this data.

We open the research topic with a review of antibody repertoires in fish by Fillatreau and colleagues (2). This review provides a description of the organization of fish Ig loci, with a particular emphasis on their heterogeneity between species, and presents recent data on the structure of the expressed Ig repertoire in healthy and infected fish. This is followed by a review of antibody repertoires in pigs (3). In pigs, the fetal repertoire develops without maternal influences and the precocial nature of multiple offspring provides investigators with the opportunity to study the influence of environmental and maternal factors on repertoire development.

Next, we take a closer look at the human repertoire. Vas et al. (4) discuss the role of natural antibodies (Nabs). Mostly, IgM antibodies are produced in the absence of exogenous antigen challenge. The composition of the early immune repertoire is highly enriched for NAbS, which are polyreactive and often autoreactive. Included in Nabs are antibodies that recognize damaged and senescent cells, often via oxidation-associated neo-determinants. Clinical surveys have suggested that anti-apoptotic cell (AC) IgM NAbS may modulate disease activity in some patients with autoimmune disease. This review is followed by a comparative study by Mroczek and colleagues (5) of the antibody repertoire expressed by immature, transitional, mature, memory IgD⁺, memory IgD⁻, and plasmacytes isolated from the blood of a single individual. Differences observed between the Igs produced by these cells indicate that studies designed to correlate repertoire expression with diseases of immune function will likely require deep sequencing of B cells sorted by subset. The next paper highlights secondary mechanisms of antibody diversification that act in addition to

V(D)J recombination and SHM of the complementary determining regions (CDRs) of the antibody that create the antigen-binding site (6). These secondary mechanisms include V(DD)J recombination (or D-D fusion), SHM-associated insertions and deletions, and affinity maturation and antigen contact by non-CDR regions of the antibody. Next is an analysis of age-related changes in the antibody repertoire following vaccination by Wu et al. (7). Clustering analysis of high-throughput sequencing data enables us to visualize the response in terms of expansions of clonotypes, changes in CDR-H3 characteristics, and SHM as well as identifying the commonly used IgH genes. This study highlights a number of areas for future consideration in vaccine studies of the elderly.

Finlay and Almagro (8) pull all of these strands together in the final research based article, which reviews the structural studies and fundamental principles that define how antibodies interact with diverse targets. They compare the antibody repertoires and affinity maturation mechanisms of humans, mice, and chickens, as well as the use of novel single-domain antibodies in camelids and sharks. These species utilize a plethora of evolutionary solutions to generate specific and high-affinity antibodies. The various solutions used by these species illustrate the plasticity of natural antibody repertoires. They end their article by discussing man-made antibody repertoires that have been designed and validated in the last two decades. Together, these comparative studies of natural and man-made repertoires served as tools to explore how the size, diversity, and composition of a repertoire impact the antibody discovery process.

High-throughput sequencing is tailor made for the study of antibody repertoires. However, the diversity of the sequences that is obtained from these studies is immense, and thus requires the development of new and friendly bioinformatics techniques to analyze and interpret the data. The final two articles are devoted to methods that can be used for these purposes. The first issue is quality control. Michaeli et al. (9) present a method for automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing. Their paper describes Ig high-throughput sequencing cleaner (Ig-HTS-cleaner), a program containing a simple cleaning procedure that successfully deals with pre-processing of Ig sequences derived from HTS, and Ig insertion-deletion identifier (Ig-Indel-identifier), a program for identifying legitimate and artifact insertions and/or deletions (indels). These programs were designed for analyzing Ig gene sequences obtained by 454 sequencing, but they are applicable to all types of sequences and sequencing platforms. Finally,

Rogosch et al. (10) present an easy-to-use Microsoft® Excel® based software, named immunoglobulin analysis tool (IgAT), for the summary, interrogation, and further processing of IMGT/HighV-QUEST output files. IgAT generates descriptive statistics and high-quality figures for collections of murine or human Ig heavy or light chain transcripts ranging from 1 to 150,000 sequences. In addition to traditionally studied properties of Ig transcripts – such as the usage of germline gene segments, or the length and composition of the CDR-3 region – IgAT also uses published algorithms to calculate the probability of antigen selection based on somatic mutational patterns, the average hydrophobicity of the antigen-binding sites, and predictable structural properties of the CDR-H3 loop according to Shirai's H3-rules.

The authors that contributed to this volume hope that the reader will find this research topic interesting, thought-providing, and informative. We invite you to read the following articles and immerse yourself in the fascinating world of IgS. In the near term future, this world is likely to continue to provide new venues for the diagnosis, treatment, or prevention of disease.

REFERENCES

- Hirano M, Das S, Guo P, Cooper MD. The evolution of adaptive immunity in vertebrates. *Adv Immunol* (2011) **109**:125–57. doi:10.1016/B978-0-12-387664-5.00004-2
- Fillatreau S, Six A, Magadan S, Castro R, Sunyer JO, Boudinot P. The astonishing diversity of Ig classes and B cell repertoires in teleost fish. *Front Immunol* (2013) **4**:28. doi:10.3389/fimmu.2013.00028
- Butler JE, Wertz N. The porcine antibody repertoire: variations on the textbook theme. *Front Immunol* (2012) **3**:153. doi:10.3389/fimmu.2012.00153
- Vas J, Gronwall C, Silverman GJ. Fundamental roles of the innate-like repertoire of natural antibodies in immune homeostasis. *Front Immunol* (2013) **4**:4. doi:10.3389/fimmu.2013.00004
- Mroczeck ES, Ippolito GC, Rogosch T, Hoi KH, Hwangpo TA, Brand MG, et al. Differences in the composition of the human antibody repertoire by B cell subsets in the blood. *Front Immunol* (2014) **5**:96. doi:10.3389/fimmu.2014.00096
- Briney BS, Crowe JE Jr. Secondary mechanisms of diversification in the human antibody repertoire. *Front Immunol* (2013) **4**:42. doi:10.3389/fimmu.2013.00042
- Wu YC, Kipling D, Dunn-Walters DK. Age-related changes in human peripheral blood IgM repertoire following vaccination. *Front Immunol* (2012) **3**:193. doi:10.3389/fimmu.2012.00193
- Finlay WJ, Almagro JC. Natural and man-made V-gene repertoires for antibody discovery. *Front Immunol* (2012) **3**:342. doi:10.3389/fimmu.2012.00342
- Michaeli M, Noga H, Tabibian-Keissar H, Barshack I, Mehr R. Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing. *Front Immunol* (2012) **3**:386. doi:10.3389/fimmu.2012.00386
- Rogosch T, Kerzel S, Hoi KH, Zhang Z, Maier RF, Ippolito GC, et al. Immunoglobulin analysis tool: a novel tool for the analysis of human and mouse heavy and light chain transcripts. *Front Immunol* (2012) **3**:176. doi:10.3389/fimmu.2012.00176

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 08 January 2015; accepted: 16 January 2015; published online: 05 February 2015.

*Citation: Schroeder HW Jr (2015) The evolution and development of the antibody repertoire. *Front. Immunol.* **6**:33. doi: 10.3389/fimmu.2015.00033*

*This article was submitted to B Cell Biology, a section of the journal *Frontiers in Immunology*.*

Copyright © 2015 Schroeder. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



The astonishing diversity of Ig classes and B cell repertoires in teleost fish

Simon Fillatreau¹, Adrien Six^{2,3}, Susanna Magadan⁴, Rosario Castro⁴, J. Oriol Sunyer⁵ and Pierre Boudinot^{4*}

¹ Deutsches Rheuma-Forschungszentrum, Leibniz Institute, Berlin, Germany

² UPMC Univ Paris 06, UMR 7211, "Immunology, Immunopathology, Immunotherapy," F-75013 Paris, France

³ UMR 7211, "Immunology, Immunopathology, Immunotherapy," CNRS, Paris, France

⁴ Virologie et Immunologie Moléculaires, Institut National de la Recherche Agronomique, Jouy-en-Josas, France

⁵ Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, USA

Edited by:

Harry W. Schroeder, University of Alabama at Birmingham, USA

Reviewed by:

Michael Zemlin, Philipps University Marburg, Germany

Peter D. Burrows, University of Alabama at Birmingham, USA

***Correspondence:**

Pierre Boudinot, Virologie et Immunologie Moléculaires, Institut National de la Recherche Agronomique, Domaine de Vilvert, 78352 Jouy-en-Josas, France.
e-mail: pierre.boudinot@jouy.inra.fr

With lymphoid tissue anatomy different than mammals, and diverse adaptations to all aquatic environments, fish constitute a fascinating group of vertebrate to study the biology of B cell repertoires in a comparative perspective. Fish B lymphocytes express immunoglobulin (Ig) on their surface and secrete antigen-specific antibodies in response to immune challenges. Three antibody classes have been identified in fish, namely IgM, IgD, and IgT, while IgG, IgA, and IgE are absent. IgM and IgD have been found in all fish species analyzed, and thus seem to be primordial antibody classes. IgM and IgD are normally co-expressed from the same mRNA through alternative splicing, as in mammals. Tetrameric IgM is the main antibody class found in serum. Some species of fish also have IgT, which seems to exist only in fish and is specialized in mucosal immunity. IgM/IgD and IgT are expressed by two different sub-populations of B cells. The tools available to investigate B cell responses at the cellular level in fish are limited, but the progress of fish genomics has started to unravel a rich diversity of IgH and immunoglobulin light chain locus organization, which might be related to the succession of genome remodelings that occurred during fish evolution. Moreover, the development of deep sequencing techniques has allowed the investigation of the global features of the expressed fish B cell repertoires in zebrafish and rainbow trout, in steady state or after infection. This review provides a description of the organization of fish Ig loci, with a particular emphasis on their heterogeneity between species, and presents recent data on the structure of the expressed Ig repertoire in healthy and infected fish.

Keywords: fish, antibody, repertoire, evolution, B cells

INTRODUCTION

Teleost fish form a large zoological group with about 40,000 identified species, in comparison to 10,000 species for birds, and only around 5700 species for mammals. Fish are heterogeneous with regards to size, morphology, physiology, and behavior. They are ubiquitous throughout almost all aquatic environments, which have diverse oxygen concentrations, water pressures, temperatures, and salinities. Related representatives from the same group can be found in different ecosystems. For instance, Perciformes are adapted to both freshwater and marine habitats, including Antarctic. These diverse milieus certainly host a broad variety of pathogens. Fish can be infected by viruses (rhabdoviruses, bornaviruses, reoviruses, nodaviruses, iridoviruses, herpesviruses, etc.), bacteria (*Vibrio*, *Aeromonas*, *Flavobacterium*, *Yersinia*, *Lactococcus*, *Mycobacterium*, etc.), and many parasites. Thus, it is expected that a considerable diversity of host/pathogen interactions characterize fish immune defense mechanisms.

Most of our current knowledge on the immune systems and pathogens of fish comes from aquaculture species. In this context, pathogen diagnostic and vaccination are of considerable economic

importance. As an illustration of this, the vaccination program established in Norway to protect Atlantic salmon against vibriosis and furunculosis during the last decades has dramatically reduced the impact of these pathogens, yielding a sharp increase in salmon production that now allows an export value of more than 35 billions Norwegian Kroner (close to 5 billions €) per year. The main aquaculture species of interest for immunology are rainbow trout and Atlantic salmon (*Salmo salar*, Salmoniformes), common, and crucian carp (*Cyprinus carpio* and *Carassius auratus*, Cypriniformes), channel catfish (*Ictalurus punctatus*, Siluriformes), tilapia, sea bass, and sea bream (*Oreochromis niloticus*, *Dicentrarchus labrax*, and *Sparus aurata*, Perciformes), Japanese flounder (*Paralichthys olivaceus*, Pleuronectiformes), as well as cod (*Gadus morhua*, Gadiformes). The immune systems of several additional species of economical importance in Asia like Grass carp (*Ctenopharyngodon idella*, Cypriniformes), and mandarin fish (*Siniperca chuatsi*, Perciformes) have been increasingly studied during the last years. In addition, a few freshwater fish species originally studied in developmental biology for their capacity to provide eggs, or for their ecological/morphological characteristics, later became experimental models in Immunology. These include

zebrafish (*Danio rerio*, Cypriniformes), medaka (*Oryzias latipes*, Beloniformes/Cyprinodontiformes), and stickleback (*Gasterosteus aculeatus*, Gasterosteiformes). In sum, it stands out that our knowledge of fish immunology relates only to a minor fraction of the 40,000 known fish species. It is therefore important not to generalize observations made in individual groups, especially since our knowledge on the model species listed above already illustrates that the organization of the immune system differs among distinct fish species.

Besides its direct relevance for aquaculture, the study of the immune system of fish is also of interest to understand the evolution of the adaptive immune system in Vertebrates. The primordial adaptive immune system of extinct vertebrates is not accessible, but it can be inferred through comparative analyses of the B and T cell systems from distant living groups like fish and mammals. Although fish lack bone marrow and lymph nodes, fish infections by bacterial or viral pathogens can lead to the production of specific antibodies, which in some cases correlates perfectly with protection against re-infection by these pathogens. Such a protection may persist for more than 1 year. It is therefore possible to compare how the humoral immune system functions in fish and in mammals.

Research on the immune system of fish has generally been limited by the lack of reagents suitable for classical cellular immunology research, but it has greatly benefited from the sequencing of their genomes (Table 1), which have particular structural features directly relevant for their immune system. In particular, a cycle of tetraploidization and re-diploidization occurred during the early evolution of fish genomes, which was followed by further cycles of whole-genome duplications, and differential loss of various genome parts during the subsequent evolution of many fish families (Figure 1). As a result, fish genomes are especially heterogeneous. Some genes involved in the immune system have been affected by these re-modelings; in fact, the great number of gene duplicates has probably played an important role in the diversification of the immune genes through sub-functionalization and specific adaptations. This might also account for the fact that

the immunoglobulin (Ig) loci of some fish species are among the largest and most complex described yet. Salmonids have two IgH loci per haplotype with several hundreds of V genes, while mammals have only one IgH loci per haplotype and fewer VH genes.

The availability of genomic resources has been particularly useful to investigate B cell repertoires in fish, both for the description of the genomic organization of Ig loci, which defines the potential repertoire, and for the characterization of the primary repertoire expressed by B cells in healthy and infected fish (Jerne, 1971). When considering the importance of efficient adaptive immune responses for the control of infectious diseases, and for successful vaccination, one realizes the relevance of understanding how lymphocyte repertoires are selected during B cell development and modified upon antigenic challenge. In this review, we will first examine fish Ig classes, the structure of the loci, and the IgH splicing patterns. We will then study the B cell system and the features of the available (expressed) repertoires of antibodies in healthy or infected fish.

DIVERSIFICATION OF IG GENES IN FISH: POTENTIAL REPERTOIRES AND DIVERSIFICATION MECHANISMS

Ig LOCI IN FISH

Fish have three Ig classes

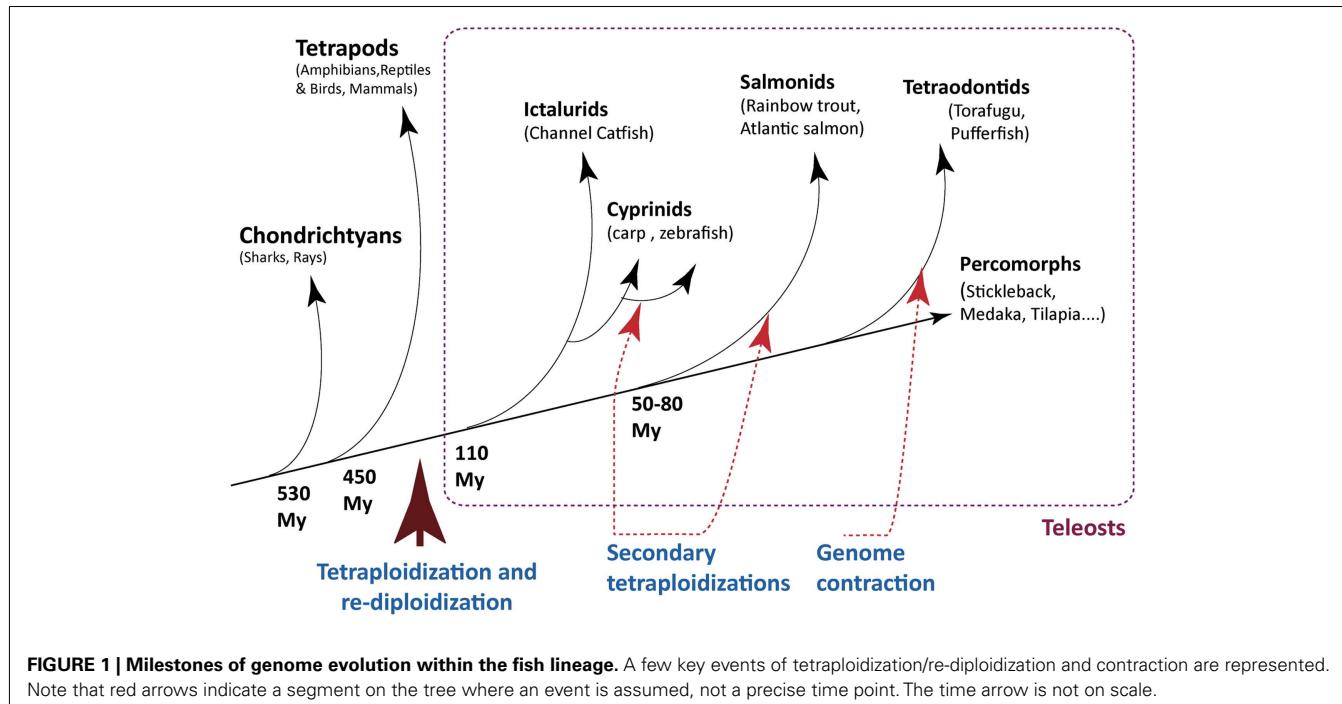
Three classes of Ig have been identified in teleost fish. These are IgM, which is found in all vertebrate species (reviewed in Flajnik and Kasahara, 2009), IgD, which also has a wide distribution among vertebrates, and IgT/Z (for Teleost/Zebrafish), which is specific to fish. Hereafter, fish IgM, D, and T/Z classes refer to the protein products of the isotypes μ , δ , and τ/ζ , respectively, which correspond to their associated constant genes.

IgM was the first Ig class identified in fish. It can be expressed at the surface of B cells or secreted. Secreted tetrameric IgM represents the main serum Ig in fish.

IgD was initially thought to be expressed only in rodents and primates, and to be of recent evolutionary origin. However, the first fish IgD was identified in Wilson et al. (1997) in the channel catfish.

Table 1 | Status of genome sequencing of the main model species for fish immunology.

AQUACULTURE SPECIES	
Rainbow trout (<i>Oncorhynchus mykiss</i>)	Genome in progress
Atlantic salmon (<i>Salmo salar</i>)	Genome in progress
Atlantic cod (<i>Gadus morhua</i>)	Genome published (Star et al., 2011)
Common carp (<i>Cyprinus carpio</i>)	Genome published (Henkel et al., 2012)
Crucian carp (<i>Carassius carassius</i>)	
Channel catfish (<i>Ictalurus punctatus</i>)	Genome in progress
Tilapia (<i>Oreochromis niloticus</i>)	Genome available at http://www.ensembl.org/Oreochromis_niloticus/Info/Index
Sea bass (<i>Dicentrarchus labrax</i>)	Genome in progress
Sea bream (<i>Sparus aurata</i>)	
Japanese flounder (<i>Paralichthys olivaceus</i>)	
MODEL SPECIES	
Zebrafish (<i>Danio rerio</i>)	Genome available at http://www.ensembl.org/Danio rerio/Info/Index
Medaka (<i>Oryzias latipes</i>)	Genome published (Kasahara et al., 2007)
Three-spined stickleback (<i>Gasterosteus aculeatus</i>)	Genome published (Jones et al., 2012)



It differs from mammalian IgD because it is a chimeric protein containing a C μ 1 domain followed by a number of C δ . This chimeric structure was also found in Atlantic salmon (Hordvik et al., 1999), and other fish species (Stenvik and Jørgensen, 2000; Aparicio et al., 2002; Hordvik, 2002; Srisapoome et al., 2004; Xiao et al., 2010). To date, no complete fish IgD heavy chain without C μ 1 has been described. Intriguingly, a similar C μ 1–C δ structure has been discovered in some non-fish species of the order of the Artiodactyls (Zhao et al., 2002, 2003). Fish IgD also differs from eutherian IgD by the large number (7–17) of C δ domains it can contain, and by the absence of a hinge. Secreted IgD have been found in catfish (Edholm et al., 2010), and in rainbow trout (Ramirez-Gomez et al., 2012), but with some differences because it did not contain V domain in the former, while it did in rainbow trout. Of note, IgD has been found in most vertebrates, and it has orthologs even in Chondrichthyans (known as IgW), suggesting that it represents a primordial Ig class, like IgM (Ohta and Flajnik, 2006). To date, IgD seems to be missing only in birds, and in few mammalian species. No IgD sequence was found in the chicken IgH locus (Zhao et al., 2000) and seems to be absent from the chicken genome. IgD could not be found from available sequences from duck and ostrich either (Lundqvist et al., 2001; Huang et al., 2012). In the same line, IgD is apparently absent from the elephant and opossum IgH loci (Wang et al., 2009; Guo et al., 2011).

IgT/IgZ was discovered in Hansen et al. (2005) in rainbow trout (IgT) and zebrafish (IgZ; Danilova et al., 2005). It does not exist in other vertebrates but fish. IgH τ / ζ may contain different numbers of C domains: four C domains are found in most species (Salinas et al., 2011), whereas stickleback (*G. aculeatus*) has three and fugu (*Takifugu rubripes*) has two. In carp (*C. carpio*) IgT is a chimeric protein containing a C μ 1 domain and a C τ / ζ domain (Savan et al., 2005). No Ig τ / ζ locus could be found in the Medaka genome or in

the Channel catfish, but it might be identified in catfish when the full genome sequence will be available. Recent studies performed in trout demonstrate that IgT is especially critical for the protection of mucosal territories in this species (Zhang et al., 2010), as suggested by the fact that the local ratio of IgT to IgM is >60-fold higher in the gut mucus than in serum. Furthermore, fish surviving an infection by the gut parasite *Ceratomyxa shasta* had elevated titers of parasite-specific IgT only in the gut mucus but not in the serum, while high titers of parasite-specific IgM were measured in the serum but generally not in the mucus. Additionally, as for IgA in human, an important property of IgT in the gut of rainbow trout seems to be its ability to recognize and coat a large percentage of luminal bacteria at steady state. Secreted IgT is found in trout serum as a monomer, and in mucus as a tetramer (Zhang et al., 2010).

Remarkably, neither IgG nor IgE are present in fish, even though long-lasting protection against secondary infection exists, and many parasites can infect fish.

Fish IgH loci: structure and number across fish species

The archetypal structure of the IgH loci follows a pattern of translocon organization with a region containing VH genes in 5', followed by units comprising several D, J, and then C region genes in 3'. The D τ -J τ -C τ cluster(s) encoding IgT specific genes are generally located between the region containing the VH genes and the D μ / δ -J μ / δ -C μ -C δ locus. This structure is found for example in the zebrafish, grass carp, and fugu (Figure 2A). In this case, the configuration of IgH loci imposes the alternative production of either IgT or IgM/D rearrangements at a given locus since the recombination of VH to D μ deletes the D τ -J τ -C τ region(s). Since most VH genes are located upstream of both DH τ and D μ / δ , they can probably be used by IgT, IgM, and IgD

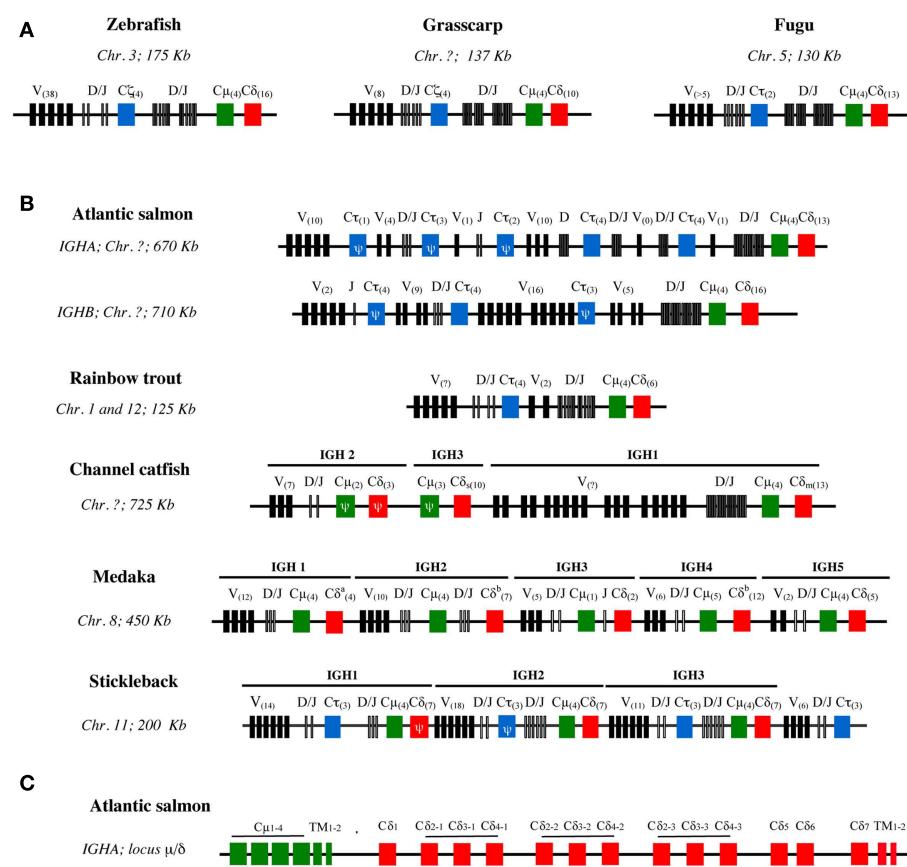


FIGURE 2 | Schematic structure of IgH loci in different teleost species.

(A) IgH loci with archetypic structure in zebrafish, grass carp, and fugu. **(B)** Variants of IgH structure found in other species with partial or complete duplications present in different chromosomes (Chr.) (Atlantic salmon, rainbow trout) or in the same chromosome (channel catfish, three-spined stickleback, and Japanese medaka) (Chr.). The schemes are not in scale and depict the genomic configuration of V sets (black boxes), D and J sets (narrow gray boxes), and CH gene sets. C μ are represented as green boxes, C δ as red

boxes, and C τ/ζ as blue boxes. The number of in frame V genes and CH exons are indicated in brackets within boxes. CH sequences with frameshift mutations are considered as pseudogenes (Ψ). Catfish IgH: C δ_s and C δ_m correspond to the secreted and membrane IgD coding genes, respectively. Medaka IgH: in the C δ^a , the genomic sequence presents a gap and the actual number of C δ domains is unknown; C δ^b indicates the presence of C μ domains inserted between C δ exons. The "?" symbol indicates a lack of data. **(C)** Detailed exon structure of the IgHA μ - δ region in Atlantic salmon.

(Danilova et al., 2005; Hansen et al., 2005). A large number of VH genes are either pseudogenes, or their sequence is not complete in the genome assembly. Therefore, the diversity of functional VH genes is difficult to estimate. Beyond these general features, the structure of the loci coding for the isotypes corresponding to IgM, IgD, and IgT is surprisingly diverse among teleost fish species, due to successive episodes of genome duplications and gene loss.

Various number of IgH loci can be found in teleost species. The number of IgH loci varies among teleosts, and in some cases isoloci can even be found on different chromosomes (Figure 2B).

Salmonids such as Atlantic Salmon and rainbow trout possess two IgH isoloci (IgHA and IgHB) due to the tetraploidization of Salmonidae (Yasuike et al., 2010). The two corresponding IgM subtypes seem to be expressed at the mRNA level in Atlantic salmon and brown trout, but only one is found in rainbow trout and arctic

char, suggesting that one of the two isoloci may be non-functional in these last two species. In Atlantic salmon, considering both IgHA and IgHB isoloci, there are eight C τ loci with variable numbers of D τ and J τ genes likely due to tandem duplications, but only three out of these eight loci seem to be functional (two for IgHA and one for IgHB). In contrast, there is only one D μ/δ -J μ/δ -C μ -C δ region per isolocus.

Cyprinids can also have different types of IgH loci. Zebrafish has only one IgH locus with the archetypic structure, as mentioned above (Danilova et al., 2005). The common carp has two subclasses of IgT/Z: IgZ1 is similar to the zebrafish IgZ while the IgZ2 contains a C μ 1 domain (Ryo et al., 2010). It seems that the two carp IgZ are expressed from two distinct loci, but it is not clear at present whether these loci are located on the same chromosome. The common carp genome has been recently sequenced, and may provide novel information when fully annotated (Henkel et al., 2012).

In other species like channel catfish, medaka, and three-spined stickleback, tandem duplications of the IgH locus have been found (**Figure 2B**). The channel catfish IgH region contains three μ/δ loci, yet only 1 μ is functional and τ/ζ has not been found so far. The absence of IgT, which still has to be confirmed by full genome sequencing, might be due to a gene loss in the early evolution of Ictalurids. Intriguingly, in catfish the membrane IgD and the (V-less) secreted IgD are always produced from the two different functional C δ (Bengtén et al., 2006). It remains to be determined whether they could be expressed from the same haplotype. In the medaka genome, five regions encoding constant domains of IgM and IgD have been identified in one large locus (Magadán-Mompó et al., 2011). The analysis of Expressed Sequence Tags (ESTs) suggests that the IGH3 region is disorganized and might be non-functional (**Figure 2B**). No IgT gene has been found so far in this species. In the stickleback genome, three sets of $\tau/\zeta-\mu-\delta$ loci separated by VH-containing regions have been described, evoking recombination units as found in mouse λ light chains or shark IgH loci (Bao et al., 2010; Gambón-Deza et al., 2010).

The structure of the IgH δ locus differs between fish species.

A precise examination of fish IgH shows that the structure of IGH δ is remarkably heterogeneous among fish species with frequent C-domain duplications, while IgH μ and likely IgH τ appear to be more conserved. For example, C δ 2–C δ 3–C δ 4 domains are repeated three times in Atlantic salmon IgHA (**Figure 2C**) and catfish, and four times in zebrafish and Atlantic salmon IgHB. In puffer fish, the IgD gene comprises a longer tandem C δ 1 → C δ 6 duplication (Saha et al., 2004). The rainbow trout IgD gene is also particular as it carries a C δ 1–C δ 2a–C δ 3a–C δ 4a–C δ 2b–C δ 7 configuration, which seems to be the result of a first duplication of C δ 2–C δ 4 present in C δ 1–C δ 2–C δ 3–C δ 4–C δ 5–C δ 6–C δ 7, leading to C δ 1–C δ 2a–C δ 3a–C δ 4a–C δ 2b–C δ 3b–C δ 4b–C δ 5–C δ 6–C δ 7, followed by deletion of the C δ 3b–C δ 6 domains (Hansen et al., 2005). In the Japanese flounder and stickleback there is no C δ domain duplication (Hirono et al., 2003; Hansen et al., 2005; Bao et al., 2010; Gambón-Deza et al., 2010). Of note, fish IgM and IgD are co-produced through alternative splicing of a long pre-mRNA containing the VDJ region, the C μ exons, and the C δ exons, as in mammals (**Figure 3A**). Precisely, fish IgH δ mature transcripts are produced by splicing of the donor site at the end of the C μ 1 exon to the acceptor site of the first C δ exon (Wilson et al., 1997), which results in a chimeric C μ 1/C δ molecules.

Different Ig splicing patterns are used by distinct fish species to generate membrane IgM

In mice and humans, membrane, and secreted IgM H chains are produced from the same pre-mRNA through alternative splicing. A membrane Ig μ transcript is made if a cryptic splice site located within C μ 4 is spliced to the acceptor site of the transmembrane (TM)1 exon, and a secreted Ig μ transcript is produced when the mRNA is polyadenylated between the last constant (C) region domain C μ 4 and the TM exons. In fish, membrane Ig μ transcripts have the TM exons spliced to the donor site located at the 3'end of the C μ 3 exon, hence they lack the last C μ domain (C μ 4) that is present in the secreted Ig μ transcripts (**Figure 3B**;

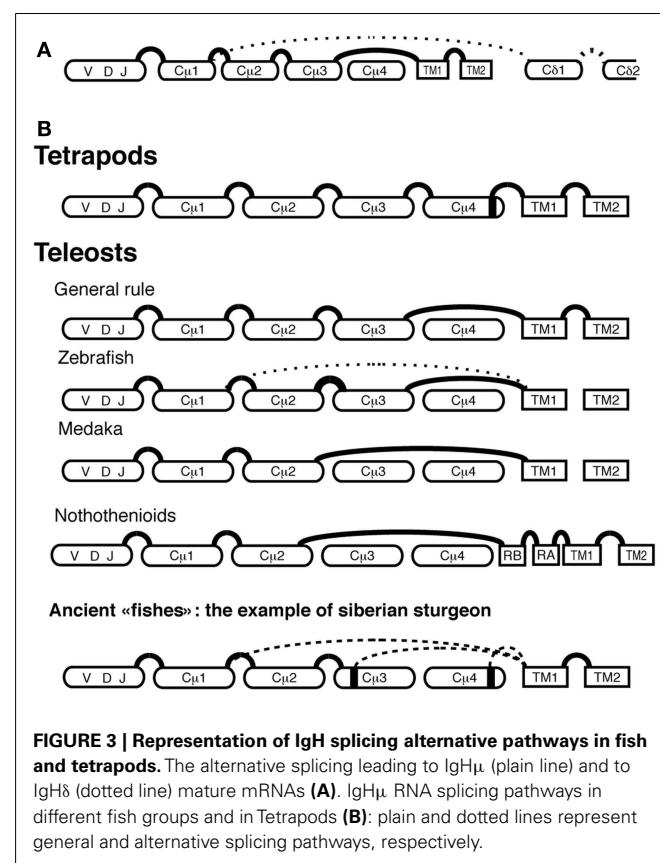


FIGURE 3 | Representation of IgH splicing alternative pathways in fish and tetrapods. The alternative splicing leading to IgH μ (plain line) and to IgH δ (dotted line) mature mRNAs (**A**). IgH μ RNA splicing pathways in different fish groups and in Tetrapods (**B**): plain and dotted lines represent general and alternative splicing pathways, respectively.

Bengtén et al., 1991; Lee et al., 1993; van Ginkel et al., 1994). Exceptions to this rule have been found in different species. The medaka membrane Ig μ lacks both C μ 3 and C μ 4 domains because the TM exons are spliced directly to the 3'end of the C μ 2 domain (Magadán-Mompó et al., 2011). In the Antarctic Notothenioids fish, membrane Ig μ transcripts also lack these domains (Coscia et al., 2010) but two exons consisting of 39-nt (RA and RB) are present between the C μ 2 and TM1 exons (Coscia et al., 2010). This splicing pattern, which is found in most of the Antarctic Notothenioids, may represent an adaptive selection of IgM during Notothenioid evolution (Coscia and Oreste, 2003). In the zebrafish, in addition to the classical VDJ–C μ 1–C μ 2–C μ 3–TM1–TM2 mRNA, an alternative VDJ–C μ 1–TM1–TM2 membrane Ig μ transcript has been reported, which encodes only one CH domain (Hu et al., 2011). This implies that B cells can express two different forms of membrane IgM in this species, which increases the number of expressed Ig isotypes. Noteworthy, the functional implication of these various splicing patterns for B cell functions is unknown.

In ancient lineages of fish such as holosteans, the bowfin (*Amia calva*), and the long-nose gar (*Lepisosteus osseus*) a remarkable diversity of splicing patterns of the membrane Ig μ transcripts was also observed (Wilson et al., 1995a,b). In chondrosteans, another ancient fish lineage, the diversity of membrane Ig μ transcripts is even higher: in Siberian sturgeon (*Acipenser baerii*) the TM1 exon is alternatively spliced to three possible donor sites: a cryptic site at the end of C μ 4, a cryptic site at the end of C μ 3, and the

donor splice site at the 3' end of C μ 1, leading to IgM H chains with four, two, or only one complete C μ domain(s) (Lundqvist et al., 2009; **Figure 3B**). The shortest membrane Ig μ splice variant might have specialized functions because it was retrieved only in transcripts expressing VH2 (Lundqvist et al., 2009). This diversification of splicing pathways to produce membrane IgM in the “ancient fish” lineages evokes a highly diverse situation after whole-genome duplication in the early fish evolution, followed by standardization to the C μ 3 → TM1 splicing pattern in teleosts before their great radiation. However, the particular situations found in ice fish (Notothenioids) or even in zebrafish or medaka indicate that this standardization is not universal.

IgL loci in teleost fish

Four immunoglobulin light chain (IgL) isotypes have been described in teleosts: L1, L2, L3, and λ (Edholm et al., 2009; Bao et al., 2010). A recent comprehensive phylogenetic analysis of vertebrate VL and CL sequences suggested that fish L1 and L3 chains are κ orthologous (Criscitiello and Flajnik, 2007), and fish L2 are orthologs of Xenopus σ (Partula et al., 1996). The current general classification of IgL from all vertebrates distinguishes four clans based on phylogenetic relationships: κ (mammalian κ , elasmobranch Type III, Teleost L1, and L3, Xenopus ρ), λ (mammalian λ , elasmobranch type II), σ (Xenopus σ , teleost L2, elasmobranch type IV), and σ cart (σ -cart, that is restricted to elasmobranch).

Light chain genes in fish genomes are found as multiple VL–JL–CL units. The genomic organization of VL–JL–CL unit is conserved in teleosts. For the L1 and L3 loci, the V genes are in opposite transcriptional orientation with respect to the J and C segments. In contrast, in L2 and λ clusters V, J, and C genes are in same orientation (Daggfeldt et al., 1993; Ghaffari and Lobb, 1993, 1997; Timmus et al., 2000), with the exception of stickleback L2 (Bao et al., 2010). In a genome wide study in zebrafish, such clusters have been found in five different chromosomes (Zimmerman et al., 2008). Interestingly, VL–JL rearrangements between distinct units were reported in this species, which might be a means of increasing the potential combinatorial diversity.

It is intriguing that to date no pseudo light chain corresponding to eutherian mammal VpreB or λ 5 has been reported in fish. In mammals, these chains play a crucial role in the stepwise process of Ig chains rearrangements that take place during B cell development. A deficiency in VpreB or λ 5 results in a block of B cell development at the pre-B cell stage in mice (Kitamura et al., 1992). In fish, it is still unknown if an alternative pre-B cell receptor that lacks the VpreB/ λ 5 surrogate light chain forms during B cell development. In fact, it is not known if Ig gene rearrangements in fish follow the ordered model described in mouse and human; moreover, no pre-T α receptor has been found in fish, while it was recently discovered in sauropsids (Smelty et al., 2010). Similarly, the mechanisms ensuring allelic exclusion in fish are unknown.

PATHWAYS AND ENZYMATIC MACHINERY OF Ig REARRANGEMENT AND DIVERSIFICATION

The enzymatic machinery of Ig gene rearrangement: similarities between fish and mammals

The rearrangement of VDJ genes is mediated in mammals by a complex enzymatic machinery that includes recombination

activating genes (RAG)-1 and 2, proteins from the non-homologous end joining (NHEJ) pathway of repair of DNA double strand breaks, and DNA polymerases of the X family polymerase λ , polymerase μ , and terminal deoxynucleotidyl transferase (TdT). RAG are lymphocyte-specific enzymes that mediate the first steps of VDJ recombination including recognition of the Recognition Sequence Signal (RSS) situated on the sides of the Ig gene segments recruited in the rearrangement, cleavage of DNA at these RSS sites, and hairpin formation as well as resolution. The NHEJ components (Ku70, Ku80, DNAPK, XRCC4, ligase IV, and ARTEMIS) constitute the major pathway involved in the repair of the double strand DNA breaks introduced by the RAG enzymes. The resolution of the DNA breaks is preceded by the action of polymerases λ and μ , which mediate DNA deletional trimming at the junction site, and TdT, which adds “N” nucleotides in a template-independent manner in VDJ junctions. The enzymes implicated in the molecular machinery of Ig rearrangements are remarkably conserved between mammals and fish (**Table 2**).

RAG1 and RAG2 from fish were first cloned in rainbow trout (Hansen and Kaattari, 1996; Hansen, 1997) and zebrafish (Greenhalgh and Steiner, 1995; Willett et al., 1997). They are expressed in tissues where rearrangement activity is expected, and a zebrafish with a truncated RAG1, identified by screening of N-ethyl-N-nitrosourea mutants, is unable to make VDJ rearrangements, indicating that this enzyme is required for this process in zebrafish as in mammals (Wienholds et al., 2002). In line with this, V, D, and J segments of fish IgH and IgL are flanked by typical RSS (Ghaffari and Lobb, 1997; Hayman and Lobb, 2000). Of note, RSS-like heptamers and nonamers were found within some JL–CL introns (Ghaffari and Lobb, 1997) as well as in 3' region of the majority of the zebrafish CL genes (Zimmerman et al., 2011), evoking the isolated RSS heptamer recombination element located in mouse J κ –C κ intron, which can recombine with the κ -deleting element located downstream of C κ exon to delete the C κ exon and silence the Ig κ locus (Vela et al., 2008). Such process of locus inactivation might provide a mechanism to achieve allelic exclusion for fish IgL (Vela et al., 2008).

The genes coding for the main enzymes of the NHEJ machinery appear to be present in fish genomes, with (recent) duplications for some of them in zebrafish (**Table 2**). An ortholog of Ku70 was identified in zebrafish that was critical for protection from radiation-induced DNA damage because embryos in which this gene was knocked-down were highly sensitive to ionizing radiation (Bladen et al., 2007).

Orthologs of the X family of DNA polymerases involved in diversification of VDJ junctions have also been identified in fish. The gene coding for TdT was found in rainbow trout and zebrafish genomes (Hansen, 1997; Beetz et al., 2007). It is expressed in lymphoid tissues where rearrangements occur (thymus, pronephros, mesonephros, spleen, and gut). Both TCR and Ig junctions contain N diversity, suggesting that fish TdT has similar functions as in mammals. In zebrafish polymerase μ is expressed also in primary lymphoid tissues, as well as in ovary and testis (Beetz et al., 2007). Thus, the mechanisms of Ig rearrangement might be similar in teleosts and mammals.

Table 2 | Genes of the key participants of the rearrangement machinery in fish.

Reference	Zebrafish	Stickleback
<i>Rag1</i>	ENSDARG00000052122	ENSGACG00000011465
<i>Rag2</i>	ENSDARG00000052121	ENSGACG00000011461
NHEJ		
<i>Ku70</i>	ENSDARG00000090718 ENSDARG00000071551	ENSGACG00000004868
<i>Ku80</i>	ENSDARG00000068862 ENSDARG00000015599	ENSGACG00000006130
<i>XRCC4</i>	ENSDARG00000010732	? (But present in a number of other fish)
<i>DNAPK</i>	ENSDARG00000075083	ENSGACG00000001974
<i>Artemis/DCLRE1C</i>	ENSDARG00000045704	ENSGACG00000020073
<i>Ligase IV</i>	ENSDARG00000060620	ENSGACG00000014135
POLYMERASES X		
<i>Polymerases λ</i>	ENSDARG00000039613	ENSGACG00000018272
<i>Polymerases μ</i>	ENSDARG00000042507	ENSGACG00000001887
<i>TdT</i>	ENSDARG00000038540	ENSGACG00000002880

It is interesting to note that the genes coding for some of these enzymes are present in the genomes of ancient fish such as cartilaginous elasmobranch (which include shark, ray, and skates). TdT from elasmobranch has structural similarities with the mouse TdT, in agreement with the fact that both enzymes have template-independent mode of DNA elongation without strong nucleotide bias (Bartl et al., 2003). These data suggest that TdT and other polymerases from the ancient family of polymerases X were used by the rearrangement machinery even before the divergence of fish and mammals (Beetz et al., 2007).

Mechanisms of hypermutation: presence and limits

The affinity maturation of antibody responses is less efficient in cold blood vertebrates compared to mammals (Wilson et al., 1992). For example after immunization of rainbow trout with the hapten-carrier antigen TNP-KLH (trinitrophenyl-linked to keyhole limpet hemocyanin), the affinity of the antigen-specific antibody response progressively increased over 27 weeks, with initial production of low affinity antibodies, which were replaced within 5 weeks by antibodies of intermediate affinity, and after 15 weeks by antibodies that had the highest affinity for antigen (Ye et al., 2011). It is assumed that the low efficiency of the affinity maturation of the antigen-specific antibody response in fish is due to the absence of typical germinal centers (GC), which are the specialized anatomical structures supporting the selection of B cells expressing high affinity B cell receptor (BCR) for antigen in mammals (Wilson et al., 1992). However, clusters of cells containing melano-macrophages were found in spleen and kidney of channel catfish, which might represent primordial GC because activation-induced deaminase (AID) was expressed in these structures (Saunders et al., 2010). AID is a critical enzyme for somatic hypermutation and class switch recombination of Ig genes in mammals. Fish AID differ from their mammalian counterparts at the level of the catalytic sites, but puffer fish and zebrafish AID could nonetheless mediate Ig class switch recombination in

mouse B cells (Barreto et al., 2005). In catfish hypermutated IgH sites show an accumulation of G → A and C → T substitutions consistent with AID activity. However, the pattern of Ig somatic hypermutation has particular characteristics in fish, with sequence motifs containing hypermutation hotspots different from those known in mammals (Yang et al., 2006). Interestingly, there was no difference in the ratio of replacement-to-silent mutations in the complementarity determining regions (CDR), which correspond to the Ig parts involved in antigen binding, and in the framework regions, which are normally not involved in antigen binding. Thus, the mechanism of Ig somatic mutation did not co-evolve in fish with the pathways mediating selection of B cells with non-synonymous substitutions specifically within CDR-encoded regions. Fish Ig structure suggests that as in mammals CDR are most important for antigen binding, and that they form the main part of the antigen binding surface. A possible explanation for this finding is that mutated Ig sequences do not undergo positive selection for affinity maturation efficiently due to the lack of an appropriate micro-environment. In this context, the primary role of the process of somatic hypermutation might have been to diversify the available repertoire by targeting hotspot motifs preferentially located within CDR-encoded regions. Whether part of this diversity might have deleterious specificity and require particular negative selection remains unknown. In zebrafish, a comprehensive analysis of IgHμ transcripts via deep sequencing indicated that the frequency of Ig sequences with high numbers of somatic mutations increased with age (up through 1 year), in agreement with the notion that hypermutation brings a significant contribution to the diversification of the Ig repertoire (Jiang et al., 2011). Fish Ig light chains can also be subjected to hypermutation (Marianes and Zimmerman, 2011), as previously observed in shark (Lee et al., 2002). It is so far unknown whether fish AID, like mammalian ones, can specifically target additional genes with frequent translocations in tumors, repetitive sequences, and histone H3K4 trimethylation (Kato et al., 2011). The gene *aid* is found

in the genome of the main fish model species within conserved synteny groups, indicating they represent true orthologs of the mammalian gene (see zebrafish ENSDARG00000015734, stickleback ENSGACG00000010521, fugu ENSTRUG00000007079 in the Ensembl website).

THE CENTRAL B CELL SYSTEM IN FISH

Three modes of early hematopoiesis have been described in fish (Zapata et al., 2006): hematopoiesis can start in the yolk sac blood islands like in the angelfish, or in intraembryonic intermediate cell mass (ICM) as in zebrafish; alternatively it may initiate for a short time in the yolk sac before continuing in the ICM as in rainbow trout. In zebrafish, the hematopoietic activity appears at 4 days post-fertilization (dpf), but gives rise first to erythroblasts and myeloid cells. Fish B cell lymphopoiesis appears and occurs mainly in the kidney. The expression of zebrafish RAG2 was observed at 8 dpf in the pronephros of *Rag2-Gfp* transgenic fish, which was the earliest extrathymic site of RAG expression (Trede et al., 2004). AID mRNA was even detected at 2 dpf in this species by analysis of gene expression on the whole embryo (Trede et al., 2004). The first VHDHJH rearrangements were detected around 4 dpf (Danilova and Steiner, 2002), but cells expressing IgM (Lam et al., 2004) appeared in the kidney only at 3 weeks post-fertilization, suggesting a slow process of B cell maturation. In the rainbow trout, RAG expression occurred earlier, from 10 dpf onward, and membrane IgM-expressing cells became detectable at hatching (Razquin et al., 1990), around 3 wpf (Hansen, 1997). The spleen seems to have much less importance for B cell lymphopoiesis than the kidney tissue, if any.

In adult fish, B cells reside in the anterior and posterior kidney, spleen, gut lamina propria, and blood (Rombout et al., 1993; Abelli et al., 1997). Several B cell subsets can be distinguished according to their expression of distinct Ig class combinations. In some fish species two subsets of B cells can be identified by their expression of both IgM and D, or IgT only. The development of IgM⁺IgD⁺ B cells and IgT⁺ B cells involves two different pathways because in zebrafish with a deficiency in *Ikaros* gene IgT⁺ B cells are totally lacking, while IgM⁺ B cells are present, even though their appearance shows a delayed kinetic (Schorpp et al., 2006). Moreover, these two types of B cells are differently localized in the organism. IgM⁺ B cells are the main B cell population (75–80%) in spleen, kidney, and blood, while IgT⁺ B cells represent the main B cell subset (55%) in gut-associated lymphoid tissues (Zhang et al., 2010). The existence and importance of IgM⁻IgD⁺IgT⁻ B cells in fish is a matter of debate. While in most fish species it is considered that IgD is always co-expressed with IgM, a distinct population of IgM⁻IgD⁺ B cells has recently been identified in the channel catfish, which preferentially expresses σ IgL (Edholm et al., 2010). The frequency of this population is highly variable between individuals, ranging from a few percent to more than 70% of B cells within peripheral blood leukocytes. The participation of these cells to immune responses is not known.

Fish B cells show different homing patterns depending on their development and activation stages. B cell progenitors and plasma cells are dominant in the anterior kidney, while mature B cells and plasma blasts are primarily found in posterior kidney (Zwollo et al., 2005; Zwollo, 2011). Spleen leukocytes also contain B cells

that can differentiate into plasma cells. Based on these data, it can be envisioned that B cell development occurs in the anterior kidney, from where mature B cells enter the blood/lymph to reach the spleen and posterior kidney, where they can become activated and differentiate into plasma blasts and then plasma cells, which migrate back to the anterior kidney where they might subsist as long-lived cells in particular niches. Such model suggests that B cells use the same tissue for their development as plasma cells for their residence, as previously observed in mammals.

The B cell repertoire in the healthy fish

The modalities of B cell selection to produce a naïve repertoire remain unknown in fish. The development of high-throughput sequencing methods now makes possible a comprehensive description of expressed immune repertoires. The first exhaustive sequencing of a B cell expressed diversity in a vertebrate was performed in zebrafish by Weinstein et al. (2009) using 454 GS FLX pyrosequencing. In this study, whole-fish mRNA was prepared from 14 individuals belonging to 4 families, and the variable domain (VDJ) region of IgHμ sequenced. The expressed IgM repertoire was studied in quiescent state, from healthy fish that had been raised in classical aquarium environment and possessed a normal gut microbial flora. It was estimated that a large proportion of the possible V/J combinations (50–86%) were expressed. Interestingly, the distribution of VDJ diversity was similar between individuals, and identical μ heavy chains were found in distinct fish more often than expected. This study established that the expressed IgM repertoire of different fish belonging to distinct families shared some patterns, a property which was called stereotypy. The same laboratory also followed the evolution of the expressed IgM repertoire during zebrafish development (Jiang et al., 2011). In 2-weeks-old fish, the repertoire of VDJ combinations showed a high level of stereotypy, suggesting that the primary repertoire was strongly constrained. In such young fish, which have few (if any) antibody-secreting cells, the abundance, and the junctional sequence diversity of VDJ combinations correlated. In contrast, this correlation was lost in 3-month-old fish. This was likely due to the higher frequency of antibody-secreting cells in these older animals. Nevertheless, the frequencies of VDJ combinations correlated between individuals, substantiating further the notion that deterministic forces regulate the structure of the primary repertoire. The apparent contradiction between a deterministic view of the expression of VDJ combinations and the loss of correlation between VDJ frequency and diversity in adult fish may be explained by the accumulation of different numbers of plasma cells in distinct adult fish.

In a different study, a combination of CDR3 length spectratyping and pyrosequencing was used to describe the expressed IgM, IgD, and IgT repertoires in rainbow trout (Castro et al., 2013). The VDJ domains expression was studied in the spleen of naïve individuals. Clonal isogenic animals were analyzed to avoid fish-to-fish variation due to genetic heterogeneity. As in zebrafish, it was found that not all V/J combinations were expressed. In fact, only 7 out of 13 VH families were retrieved. CDR3 length spectratyping and pyrosequencing showed that spleen Ig repertoires were very diverse for all three isotypes in healthy fish. IgM and IgD repertoires were rather similar for most VH, while being significantly

different from the IgT repertoire. This observation suggested that IgM and IgD repertoires were not subjected to drastic differential selection. The strong difference between IgT and IgM/IgD CDR3 length profiles was consistent with the usage of a different set of rearrangements with specific D and J segments in B cells expressing either IgM/IgD or IgT. A more detailed analysis focused on the VDJ junctions. To compare the distributions of junctional sequences between individuals, sequence reads encoding a CDR3 region were annotated using IMGT/highV-QUEST for VH, JH, and C genes, and aggregated into “junction sequence types” (JST). The abundance distribution of JST computed from pyrosequencing datasets indicated that 90–99% of junction sequences were found less than five times, likely corresponding to naive non-expanded B cells. Only few JST were found more than 20 times, possibly reflecting the presence of few antibody-secreting cells in the spleen of these fish, in good accordance with previous studies about spleen B cell subsets in rainbow trout (Bromage et al., 2004).

Taken together, these observations indicate that all VDJ combinations are not equally expressed, and suggest that, at least in zebrafish, the expressed repertoire exhibits a significant level of stereotypy.

THE MODIFICATIONS OF FISH EXPRESSED Ab REPERTOIRES BY INFECTIONS AND VACCINES

In fish, B cell responses occur against a variety of pathogens, and must occur in microenvironments different from those described in mammals, due to the lack of GC and lymph nodes. In this context, the clonal complexity of trout B cell responses is largely unknown. The development of high-throughput sequencing approaches of Ig transcripts combined with CDR3 length spectratyping can provide comprehensive analyses of B cell responses, which are required to understand the dynamics of their clonal complexity (Ademokun et al., 2011).

Such an approach was used to characterize the B cell response of rainbow trout against a rhabdovirus, the Viral Hemorrhagic Septicemia Virus (VHSV). Clonal fish were vaccinated using an attenuated virus, then challenged 3 weeks later with the same virus, and finally analyzed after three more weeks. At this stage, all fish had neutralizing antibodies against VHSV, and increased levels of total IgM as well as IgT in serum. The titer of IgM remained more

than 10 times higher than of IgT after infection, and the ratio of IgM⁺IgT⁻/IgM⁻IgT⁺ B cells was similar between infected and control fish. CDR3 length spectratyping showed that the VHSV infection triggered a strong IgM response. Indeed, VHC μ spectratypes were extensively and significantly skewed in infected fish for all the analyzed VH, as shown by a comparison of each peak in each spectratype profile using the ISEAppeaks software (Collette and Six, 2002). Interestingly, the VH5.1-C μ profile showed a great amplification of the same peak in all infected individuals, suggesting a public response. In contrast, VHC δ profiles showed only weak and sporadic alterations, which were not statistically validated. The low contribution of IgD to the response might reflect a down regulation of its expression in activated B cells, as in human and mice. This analysis also revealed a significant IgT response in spleen of infected fish. After VHSV infection, most splenic IgT spectratyping profiles were affected, although to a lesser extent compared to IgM. No peak expansion common to all infected fish was observed for IgT, suggesting the absence of a public response. Hence, the spleen might be a site of activation for VHSV-specific IgT⁺ B cells. This is intriguing because IgT is a specialized mucosal Ig.

The molecular diversity of IgM and IgT responses was further characterized by pyrosequencing of VHC junctions for different VH groups. Since a JST corresponds to a CDR3 protein sequence associated with a (VH, JH) pair, the distribution of the relative abundance of JST in different fish provides a description of the importance of antibody clonal responses. IgM JST distributions showed that the virus induced a major shift of the IgM expressed repertoire, with appearance of a significant number of highly represented JST (Figure 4).

Further analysis indicated that these large JST sets corresponded essentially to transcripts encoding secreted IgH, hence to antibody-secreting cells. When comparing the JST expanded in different infected fish, similar VH5.1-J5 rearrangements with CDR3 of 10 amino acids were present in all individuals. The CDR3 with the amino acid sequence ARYNNNAFDY was the most frequent, but a number of other related JST were found repeated in several individuals, with exchange of small or polar amino acids: ARYNNDAFDY, ARYDDNNAFDY, ARYNSNAFDY, ARYNNVNAFDY, ARYNTNAFDY, ARYNGDAFDY, ARYSGDAFDY, and ARYNGRAFDY. Such expansion of a number

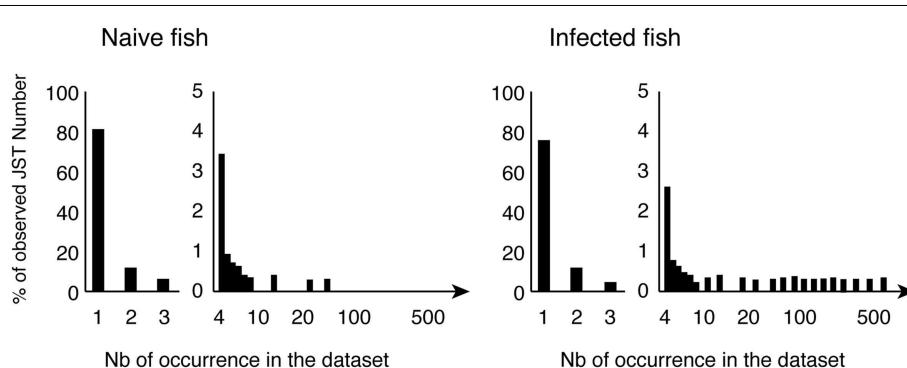


FIGURE 4 | Typical normalized distributions of JST in the pyrosequencing datasets. JST observed n times from control and virus infected fish for a given VH/C combination are represented as percentages of the total number of JST. Large clonal expansions are indicated by high number of occurrences of expressed JST in infected animals.

of similar junctions found in several fish, and differing from the most frequent one by only one (or a few) conservative substitution(s) is typical of “public” responses in mammals (Bousso et al., 1998; Lin and Welsh, 1998). Importantly, this observation suggests that rainbow trout possess a common pool of pre-existing spleen VH5.1⁺ B cells among which the public IgM response to VHSV is recruited.

This pyrosequencing study also revealed the great importance and diversity of private clonal expansions in infected fish. It is at present unclear whether these expansions represent only VHSV-specific responses or include bystander-activated cells. It will be important to clarify this point, and whether bystander effects could be beneficial or detrimental to the host. In this regard, it is intriguing that fish injected with oil-adjuvanted vaccines developed an autoimmune syndrome with autoantibodies and liver lesions (Koppang et al., 2008).

PERSPECTIVES

The aim of this review was to provide a concise description of our current knowledge of fish Ig repertoire. It is clear that a lot of important unknowns remain in fish B cell biology. As perspectives, we have listed five topics below, which might provide interesting areas for future investigation.

B CELL RECEPTOR ALLELIC EXCLUSION IN FISH

Many aspects of Ig gene rearrangement and B cell biology remain mysterious in fish. In particular, the absence of the pseudo light chains VpreB and λ 5, which are required for the formation of pre-BCR in mammals, suggests that allelic exclusion is achieved by different mechanisms in fish and mammals. In fact, the regulation of VDJ recombination to ensure allelic and isotypic exclusion in fish is far from being understood. This question evokes the situation in sharks where the IgH locus organization consists of many (up to 200) independently rearranging miniloci: in these species, the rearrangement takes place within a minilocus, and only one or few H chain genes are fully rearranged in each B cell, whereas the other loci retain their germline configuration (Malecek et al., 2008). The mechanisms by which sharks and bony fishes regulate the progression of VDJ rearrangements might reveal pathways of general interest.

MATURATION OF B CELL RESPONSE WITHOUT GC

The process of hypermutation of Ig genes observed in fish in absence of typical GC is reminiscent of the affinity maturation that can occur in mammals in extra follicular foci in the spleen red pulp (Matsumoto et al., 1996). Its potential role in the diversification of the fish Ig repertoire also bears some similarities with the fact that at least a part of human marginal zone B cell pool expresses a BCR repertoire diversified through somatic hypermutation independently of GC, even though antigen stimulation via BCR does not seem to be involved in the latter case (Weill et al., 2009). Collectively, these examples highlight the diverse utilizations made during evolution of this remarkable process of somatic hypermutation of Ig genes, for the diversification of antibody repertoires. In fish, the existence of long-term protection and antigen-specific B cell memory raises the question of differentiation of memory B cells in absence of classical GC. In fact, memory B cells expressing

high affinity, hypermutated IgG1 were found in lymphotoxin-alpha deficient mice, which lack GC (Matsumoto et al., 1996). The alternative site of memory B cells differentiation has not been identified. The modalities of memory B cell formation outside GC represent both a practical issue for vaccination and a fundamental question in B cell biology.

DIVERSITY OF B CELL REPERTOIRES

The comprehensive description of fish B cell repertoires and in-depth statistical analyses have opened the way to comparative studies of the population dynamics of B cells in different fish species. The seminal work of Quake’s group suggests that zebrafish antibody repertoires may harbor a higher level of stereotypy than expected. It will be interesting to understand if the total number of B cells present at a given time has a strong influence on such patterns: a zebrafish may contain a few millions of B cells, while a trout has around 100–1000 times more, and a large tuna probably 1000–10,000 times more. It appears likely that the constraints exerted on B cell diversity to express at once a complete repertoire able to cope properly with the diversity of relevant pathogens will be different in these species. Also, some fish species like Atlantic cod show very poor antibody responses (Espelid et al., 1991; Pilström and Petersson, 1991; Schröder et al., 1992; Magnadottir et al., 2001), when having high level of serum antibodies and a repertoire strongly skewed toward the VHIII family (Stenvik et al., 2001), possibly reflecting a particular importance of natural antibodies. These particularities must be put in the context of the absence of CD4, LI, and MHC class II molecules (hence, lack of the equivalent of a CD4⁺ T cell help activity) recently revealed by the analysis of the complete sequence of the cod genome (Star et al., 2011). As a group, teleost fish represent a rich diversity of species with a wide range of size and a complex history of whole-genome duplications. Future studies on B cell repertoires from different fish species will provide insightful information about the general rules of adaptation of this system, in fish and more generally in vertebrates.

METHODOLOGIES FOR B CELL REPERTOIRE ANALYSIS

CDR3 length spectratyping, also called Immunoscope, has been the standard technique for large-scale analysis of antigen receptors repertoire diversity for about 15 years (Pannetier et al., 1993, 1995). Systematic sequencing of “all” Ig transcripts expressed in a lymphocyte population of interest represents a step forward, and is made possible by the “next generation” sequencing technologies. A benefit of these approaches is clearly that several angles of analysis can be taken to focus on different aspects of the repertoire such as clonotypes frequency, Ig V-C or V-J CDR3 diversity, CDR3 sequence analysis, V allele identification, etc. The ability to process the complexity of the information provided in such amounts of data remains limited, and specific software developments for automatic annotation of Ig sequences, and statistical modeling of repertoire diversity can still be improved. New strategies will have to be developed, possibly from existing scoring systems. The most common is the Shannon entropy, introduced by Claude Shannon in 1948 for the information theory. Then, in 1961, Alfred Rényi has generalized the utilization of an entropy index to several functions, including Species Richness, Simpson, Quadratic, and Berger–Parker indexes to quantify the diversity, uncertainty, or randomness of a system, respectively. Among these,

Simpson's diversity and Shannon's entropy indices have already been applied to analyze TCR sequence data. A comparative review of such scoring strategies was published by Miqueu et al. (2007). Deep sequencing repertoire analysis calls for advanced statistical analysis and graphical representations, such as multivariate analysis (e.g., hierarchical clustering, principal component analysis, multidimensional scaling, etc.) and probabilistic or network modeling of sequence distributions (Mora et al., 2010; Ben-Hamo and Efroni, 2011; Murugan et al., 2012). In this perspective, different parameters can be computed to quantify the differences between repertoires at distinct levels. An important feature is the total diversity of the repertoire, which can be estimated from a dataset following approaches (Fisher et al., 1943; Efron and Thisted, 1976). At another level, a deep sequencing dataset can be summarized in various groups of sequences sharing common features (e.g., V or J gene segment, CDR3 length, sample origin, frequency), which allows comparisons between different conditions. For example, a perturbation score can be computed from the Hamming distance (Gorochov et al., 1998) to compare antibody repertoires between infected fish and a reference from control animals.

EFFECT OF TEMPERATURE ON FISH B CELL RESPONSES

While fish have colonized aquatic environments across a wide temperature range, only a few species control their internal temperature. The adaptation of fish immune system to various temperature is not fully understood, but the magnitude of the

primary response to T dependent antigens is suppressed at lower temperatures for example in the channel catfish (Bly and Clem, 1991) and carp (Le Morvan et al., 1996). More recently, it was also observed that the highest magnitude of rainbow trout specific IgM – but not IgT – response against *Yersinia ruckeri* was obtained at high temperature (25°C; Raida and Buchmann, 2007). Differential sensitivity of lymphocyte responses to temperature variations may affect immune repertoires – perhaps especially regarding natural antibodies and mucosal locations – since different pathogens may be adapted to distinct temperature ranges.

With a large number of species and a wide diversity of anatomy, physiological, and ecological adaptations to the aquatic environments and their pathogens, fish offer interesting perspectives for comparative analysis of B cell repertoire biology. New sequencing technologies have already made it possible.

ACKNOWLEDGMENTS

This work was supported by Institut National de la Recherche Agronomique and the European Community's Seventh Framework Program (FP7/2007-2013) under Grant Agreement 222719 LIFECYCLE and by the European Commission under the Work Programme 2012 of the seventh Framework Programme for Research and Technological Development of the European Union (Grant Agreement 311993 TARGETFISH). We acknowledge Dr. Oystein Evensen for helpful discussions and Dr. Vicki Lampropoulou for critical reading of the manuscript.

REFERENCES

- Abelli, L., Picchietti, S., Romano, N., Mastrolia, L., and Scapigliati, G. (1997). Immunohistochemistry of gut-associated lymphoid tissue of the sea bass *Dicentrarchus labrax* (L.). *Fish Shellfish Immunol.* 7, 235.
- Ademokun, A., Wu, Y., Martin, V., Mitra, R., Sack, U., Baxendale, H., et al. (2011). Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* 10, 922–930.
- Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-M., Dehal, P., et al. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297, 1301–1310.
- Bao, Y., Wang, T., Guo, Y., Zhao, Z., Li, N., and Zhao, Y. (2010). The immunoglobulin gene loci in the teleost *Gasterosteus aculeatus*. *Fish Shellfish Immunol.* 28, 40–48.
- Barreto, V. M., Pan-Hammarstrom, Q., Zhao, Y., Hammarstrom, L., Misulovin, Z., and Nussenzweig, M. C. (2005). AID from bony fish catalyses cell switch recombination. *J. Exp. Med.* 202, 733.
- Bartl, S., Miracle, A. L., Rumfelt, L. L., Kepler, T. B., Mochon, E., Litman, G. W., et al. (2003). Terminal deoxynucleotidyl transferases from elasmobranchs reveal structural conservation within vertebrates. *Immunogenetics* 55, 594–604.
- Beetz, S., Diekhoff, D., and Steiner, L. A. (2007). Characterization of terminal deoxynucleotidyl transferase and polymerase mu in zebrafish. *Immunogenetics* 59, 735–744.
- Bengtén, E., Leanderson, T., and Pilström, L. (1991). Immunoglobulin heavy chain cDNA from the teleost Atlantic cod (*Gadus morhua* L.): nucleotide sequences of secretory and membrane form show an unusual splicing pattern. *Eur. J. Immunol.* 21, 3027–3033.
- Bengtén, E., Quiniou, S., Hikima, J., and Waldbieser, G. (2006). Structure of the catfish IGH locus: analysis of the region including the single functional IGHM gene. *Immunogenetics* 58, 831–844.
- Ben-Hamo, R., and Efroni, S. (2011). The whole-organism heavy chain B cell repertoire from Zebrafish self-organizes into distinct network features. *BMC Syst. Biol.* 5:27. doi:10.1186/1752-0509-5-27
- Bladen, C. L., Navarre, S., Dynan, W. S., and Kozlowski, D. J. (2007). Expression of the Ku70 subunit (XRCC6) and protection from low dose ionizing radiation during zebrafish embryogenesis. *Neurosci. Lett.* 422, 97–102.
- Bly, J. E., and Clem, L. W. (1991). Temperature-mediated processes in teleost immunity: in vitro immunosuppression induced by in vivo low temperature in channel catfish. *Vet. Immunol. Immunopathol.* 28, 365–377.
- Bousso, P., Casrouge, A., Altman, J., Haury, M., Kanellopoulos, J., Abastado, J.-P., et al. (1998). Individual variations in the murine T cell response to a specific peptide reflect variability in naive repertoires. *Immunity* 9, 169–178.
- Bromage, E. S., Kaattari, I. M., Zwollo, P., and Kaattari, S. L. (2004). Plasmablast and plasma cell production and distribution in trout immune tissues. *J. Immunol.* 173, 7317–7323.
- Castro, R., Jouneau, L., Pham, H.-P., Bouchez, O., Giudicelli, V., Lefranc, M.-P., et al. (2013). Teleost fish mount complex clonal IgM and IgT responses in spleen upon systemic viral infection. *PLoS Pathog.* 9:e1003098. doi:10.1371/journal.ppat.1003098
- Collette, A., and Six, A. (2002). ISEPeaks: an Excel platform for GeneScan and Immunoscope data retrieval, management and analysis. *Bioinformatics* 18, 329–330.
- Coscia, M. R., and Oreste, U. (2003). Limited diversity of the immunoglobulin heavy chain variable domain of the emerald rockcod *Trematomus bernacchii*. *Fish Shellfish Immunol.* 14, 71–92.
- Coscia, M. R., Varriale, S., De Santi, C., Giacomelli, S., and Oreste, U. (2010). Evolution of the Antarctic teleost immunoglobulin heavy chain gene. *Mol. Phylogenet. Evol.* 55, 226–233.
- Criscitiello, M. F., and Flajnik, M. F. (2007). Four primordial immunoglobulin light chain isotypes, including lambda and kappa, identified in the most primitive living jawed vertebrates. *Eur. J. Immunol.* 37, 2683–2694.
- Daggfeldt, A., Bengtén, E., and Pilström, L. (1993). A cluster type organization of the loci of the immunoglobulin light chain in Atlantic cod (*Gadus morhua* L.) and rainbow trout (*Oncorhynchus mykiss* Walbaum) indicated by nucleotide sequences of cDNAs and hybridization analysis. *Immunogenetics* 38, 199–209.
- Danilova, N., Bussmann, J., Jekosch, K., and Steiner, L. A. (2005). The immunoglobulin heavy-chain locus in zebrafish: identification and expression of a previously unknown isotype, immunoglobulin Z. *Nat. Immunol.* 6, 295–302.
- Danilova, N., and Steiner, L. A. (2002). B cells develop in the zebrafish pancreas. *Proc. Natl. Acad. Sci. U.S.A.* 99, 13711–13716.

- Edholm, E.-S., Bengtén, E., Stafford, J. L., Sahoo, M., Taylor, E. B., Miller, N. W., et al. (2010). Identification of two IgD+ B cell populations in channel catfish, *Ictalurus punctatus*. *J. Immunol.* 185, 4082–4094.
- Edholm, E.-S., Wilson, M., Sahoo, M., Miller, N. W., Pilström, L., Wermestam, N. E., et al. (2009). Identification of Igσ and Igλ in channel catfish, *Ictalurus punctatus*, and Igλ in Atlantic cod, *Gadus morhua*. *Immunogenetics* 61, 353–370.
- Efron, B., and Thisted, R. (1976). Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika* 63, 435–447.
- Espelid, S., Rødseth, O., and Jørgensen, T. (1991). Vaccination experiments and studies of the humoral immune responses in cod, *Gadus morhua* L., to four strains of monoclonal defined *Vibrio anguillarum*. *J. Fish Dis.* 14, 185–198.
- Fisher, R. A., Steven-Corbet, A., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* 12, 42–58.
- Flajnik, M. F., and Kasahara, M. (2009). Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* 11, 47–59.
- Gambón-Deza, F., Sánchez-Espinel, C., and Magadán-Mompó, S. (2010). Presence of an unique IgT on the IGH locus in three-spined stickleback fish (*Gasterosteus aculeatus*) and the very recent generation of a repertoire of VH genes. *Dev. Comp. Immunol.* 34, 114–122.
- Ghaffari, S. H., and Lobb, C. J. (1993). Structure and genomic organization of immunoglobulin light chain in the channel catfish. An unusual genomic organizational pattern of segmental genes. *J. Immunol.* 151, 6900–6912.
- Ghaffari, S. H., and Lobb, C. J. (1997). Structure and genomic organization of a second class of immunoglobulin light chain genes in the channel catfish. *J. Immunol.* 159, 250–258.
- Gorochov, G., Neumann, A. U., Kerevur, A., Parizot, C., Li, T., Katlama, C., et al. (1998). Perturbation of CD4+ and CD8+ T-cell repertoires during progression to AIDS and regulation of the CD4+ repertoire during antiviral therapy. *Nat. Med.* 4, 215–221.
- Greenhalgh, P., and Steiner, L. A. (1995). Recombination activating gene 1 (Rag1) in zebrafish and shark. *Immunogenetics* 41, 54–55.
- Guo, Y., Bao, Y., Wang, H., Hu, X., Zhao, Z., Li, N., et al. (2011). A preliminary analysis of the immunoglobulin genes in the African elephant (*Loxodonta africana*). *PLoS ONE* 6:e16889. doi:10.1371/journal.pone.0016889
- Hansen, J., Landis, E., and Phillips, R. (2005). Discovery of a unique Ig heavy-chain isotype (IgT) in rainbow trout: implications for a distinctive B cell developmental pathway in teleost fish. *Proc. Natl. Acad. Sci. U.S.A.* 102, 6919–6924.
- Hansen, J. D. (1997). Characterization of rainbow trout terminal deoxynucleotidyl transferase structure and expression. TdT and RAG1 co-expression define the trout primary lymphoid tissues. *Immunogenetics* 46, 367–375.
- Hansen, J. D., and Kaattari, S. L. (1995). The recombination activation gene 1 (RAG1) of rainbow trout (*Oncorhynchus mykiss*): cloning, expression, and phylogenetic analysis. *Immunogenetics* 42, 188–195.
- Hansen, J. D., and Kaattari, S. L. (1996). The recombination activating gene 2 (RAG2) of the rainbow trout *Oncorhynchus mykiss*. *Immunogenetics* 44, 203–211.
- Hayman, J. R., and Lobb, C. J. (2000). Heavy chain diversity region segments of the channel catfish: structure, organization, expression and phylogenetic implications. *J. Immunol.* 164, 1916–1924.
- Henkel, C. V., Dirks, R. P., Jansen, H. J., Forlenza, M., Wiegertjes, G. F., Howe, K., et al. (2012). Comparison of the exomes of common carp (*Cyprinus carpio*) and zebrafish (*Danio rerio*). *Zebrafish* 9, 59–67.
- Hirono, I., Nam, B.-H., Enomoto, J., Uchino, K., and Aoki, T. (2003). Cloning and characterisation of a cDNA encoding Japanese flounder *Paralichthys olivaceus* IgD. *Fish Shellfish Immunol.* 15, 63–70.
- Hordvik, I. (2002). Identification of a novel immunoglobulin delta transcript and comparative analysis of the genes encoding IgD in Atlantic salmon and Atlantic halibut. *Mol. Immunol.* 39, 85–91.
- Hordvik, I., Thevarajan, J., Samdal, I., Bastani, N., and Krossøy, B. (1999). Molecular cloning and phylogenetic analysis of the Atlantic salmon immunoglobulin D gene. *Scand. J. Immunol.* 50, 202–210.
- Hu, Y.-L., Zhu, L.-Y., Xiang, L.-X., and Shao, J.-Z. (2011). Discovery of an unusual alternative splicing pathway of the immunoglobulin heavy chain in a teleost fish, *Danio rerio*. *Dev. Comp. Immunol.* 35, 253–257.
- Huang, T., Zhang, M., Wei, Z., Wang, P., Sun, Y., Hu, X., et al. (2012). Analysis of immunoglobulin transcripts in the ostrich *Struthio camelus*, a primitive avian species. *PLoS ONE* 7:e34346. doi:10.1371/journal.pone.0034346
- Jerne, N. K. (1972). "What precedes clonal selection? Ontogeny of acquired immunity," in *Proceedings of the A CIBA Foundation Symposium 1971*, Amsterdam, 1–15.
- Jiang, N., Weinstein, J. A., Penland, L., White, R. A., and Fisher, D. S. (2011). Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc. Natl. Acad. Sci. U.S.A.* 108, 5348–5353.
- Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., et al. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484, 55–61.
- Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., et al. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447, 714–719.
- Kato, L., Begum, N. A., Maxwell Burroughs, A., Doi, T., Kawaib, J., Daub, C. O., et al. (2011). Nonimmunoglobulin target loci of activation-induced cytidine deaminase (AID) share unique features with immunoglobulin genes. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2479.
- Kitamura, D., Kudo, A., Schaaf, S., Müller, W., Melchers, F., and Rajewsky, K. (1992). A critical role of lambda 5 protein in B cell development. *Cell* 69, 823–831.
- Koppang, E. O., Bjerkås, I., Haugavoll, E., Chan, E. K. L., Szabo, N. J., Ono, N., et al. (2008). Vaccination-induced systemic autoimmunity in farmed Atlantic salmon. *J. Immunol.* 181, 4807–4814.
- Lam, S. H., Chua, H. L., Gong, Z., Lam, T. J., and Sin, Y. M. (2004). Development and maturation of the immune system in zebrafish, *Danio rerio*: a gene expression profiling, in situ hybridization and immunological study. *Dev. Comp. Immunol.* 28, 9–28.
- Le Morvan, C., Deschaux, P., and Troutaud, D. (1996). Effects and mechanisms of environmental temperature on carp (*Cyprinus carpio*) anti-DNP antibody response and non-specific cytotoxic cell activity: a kinetic study. *Dev. Comp. Immunol.* 20, 331–340.
- Lee, M. A., Bengtén, E., Daggfeldt, A., Ryttig, A. S., and Pilström, L. (1993). Characterisation of rainbow trout cDNAs encoding a secreted and membrane-bound Ig heavy chain and the genomic intron upstream of the first constant exon. *Mol. Immunol.* 30, 641–648.
- Lee, S. S., Tranchina, D., Ohta, Y., Flajnik, M. F., and Hsu, E. (2002). Hypermutation in shark immunoglobulin light chain genes results in contiguous substitutions. *Immunity* 16, 571.
- Lin, M. Y., and Welsh, R. M. (1998). Stability and diversity of T cell receptor repertoire usage during lymphocytic choriomeningitis virus infection of mice. *J. Exp. Med.* 188, 1993–2005.
- Lundqvist, M., Strömberg, S., Bouchenot, C., Pilström, L., and Boudinot, P. (2009). Diverse splicing pathways of the membrane IgHM pre-mRNA in a Chondrostean, the Siberian sturgeon. *Dev. Comp. Immunol.* 33, 507–515.
- Lundqvist, M. L., Middleton, D. L., Hazard, S., and Warr, G. W. (2001). The immunoglobulin heavy chain locus of the duck. Genomic organization and expression of D, J, and C region genes. *J. Biol. Chem.* 276, 46729–46736.
- Magadán-Mompó, S., Sánchez-Espinel, C., and Gambón-Deza, F. (2011). Immunoglobulin heavy chains in medaka (*Oryzias latipes*). *BMC Evol. Biol.* 11:165. doi:10.1186/1471-2148-11-165
- Magnadottir, B., Jonsdottir, H., Helgason, S., Bjoörnsson, B., Solem, S., and Pilström, L. (2001). Immune parameters of immunised cod. *Fish Shellfish Immunol.* 10, 75–89.
- Malecek, K., Lee, V., Feng, W., Huang, J. L., Flajnik, M. F., Ohta, Y., et al. (2008). Immunoglobulin heavy chain exclusion in the shark. *PLoS Biol.* 6:e157. doi:10.1371/journal.pbio.0060157
- Marianes, A. E., and Zimmerman, A. M. (2011). Targets of somatic hypermutation within immunoglobulin light chain genes in zebrafish. *Immunology* 132, 240–255.
- Matsumoto, M., Lo, S. F., Carruthers, C. J., Min, J., Mariathasan, S., Huang, G., et al. (1996). Affinity maturation without germinal centres in lymphotoxin-alpha-deficient mice. *Nature* 382, 462–466.
- Miqueu, P., Guillet, M., Degauque, N., and Dor, J. (2007). Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. *Mol. Immunol.* 44, 1057–1064.

- Mora, T., Walczak, A. M., Bialek, W., and Callan, C. G. (2010). Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5405–5410.
- Murugan, A., Mora, T., Walczak, A. M., and Callan, C. G. (2012). Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16161–16166.
- Ohta, Y., and Flajnik, M. (2006). IgD, like IgM, is a primordial immunoglobulin class perpetuated in most jawed vertebrates. *Proc. Natl. Acad. Sci. U.S.A.* 103, 10723–10728.
- Pannetier, C., Cochet, M., Darche, S., Casrouge, A., Zöller, M., and Kourilsky, P. (1993). The sizes of the CDR3 hypervariable regions of the murine T-cell receptor beta chains vary as a function of the recombined germline segments. *Proc. Natl. Acad. Sci. U.S.A.* 90, 4319–4323.
- Pannetier, C., Even, J., and Kourilsky, P. (1995). T-cell repertoire diversity and clonal expansions in normal and clinical samples. *Immunol. Today* 16, 176–181.
- Partula, S., Schwager, J., Timmusk, S., Pilström, L., and Charlemagne, J. (1996). A second immunoglobulin light chain isotype in the rainbow trout. *Immunogenetics* 45, 44–51.
- Pilström, L., and Petersson, A. (1991). Isolation and partial characterization of immunoglobulin from cod (*Gadus morhua* L.). *Dev. Comp. Immunol.* 15, 143–152.
- Raida, M. K., and Buchmann, K. (2007). Temperature-dependent expression of immune-relevant genes in rainbow trout following *Yersinia ruckeri* vaccination. *Dis. Aquat. Org.* 77, 41–52.
- Ramirez-Gomez, F., Greene, W., Rego, K., Hansen, J. D., Costa, G., Kataria, P., et al. (2012). Discovery and characterization of secretory IgD in rainbow trout: secretory IgD is produced through a novel splicing mechanism. *J. Immunol.* 188, 1341–1349.
- Razquin, B., Castillo, A., Lopez-Fierro, P., Alvarez, F., Zapata, A., and Villena, A. (1990). Ontogeny of IgM-producing cells in the lymphoid organs of rainbow trout, *Salmo gairdneri* Richardson: an immunological and enzyme-histochemical study. *J. Fish Biol.* 36, 159.
- Rombout, J. H., Taverne-Thiele, A. J., and Villena, M. I. (1993). The gut-associated lymphoid tissue (GALT) of carp (*Cyprinus carpio* L.): an immunocytochemical analysis. *Dev. Comp. Immunol.* 17, 55–66.
- Ryo, S., Wijdeven, R. H. M., Tyagi, A., HermSEN, T., Kono, T., Karunasagar, I., et al. (2010). Common carp have two subclasses of bonyfish specific antibody IgZ showing differential expression in response to infection. *Dev. Comp. Immunol.* 34, 1183–1190.
- Saha, N. R., Suetake, H., Kikuchi, K., and Suzuki, Y. (2004). Fugu immunoglobulin D: a highly unusual gene with unprecedented duplications in its constant region. *Immunogenetics* 56, 438–447.
- Salinas, I., Zhang, Y.-A., and Sunyer, J. O. (2011). Mucosal immunoglobulins and B cells of teleost fish. *Dev. Comp. Immunol.* 35, 1346–1365.
- Saunders, H. L., Oko, A. L., Scott, A. N., Fan, C. W., and Magor, B. G. (2010). The cellular context of AID expressing cells in fish lymphoid tissues. *Dev. Comp. Immunol.* 34, 669.
- Savan, R., Aman, A., Nakao, M., Watanuki, H., and Sakai, M. (2005). Discovery of a novel immunoglobulin heavy chain gene chimera from common carp (*Cyprinus carpio* L.). *Immunogenetics* 57, 458–463.
- Schorpp, M., Bialecki, M., Diekhoff, D., Walderich, B., Odenthal, J., Maischein, H.-M., et al. (2006). Conserved functions of Ikaros in vertebrate lymphocyte development: genetic evidence for distinct larval and adult phases of T cell development and two lineages of B cells in zebrafish. *J. Immunol.* 177, 2463–2476.
- Schrøder, M., Espelid, S., and Jørgensen, T. (1992). Two serotypes of *Vibrio salmonicida* isolated from diseased cod (*Gadus morhua* L.); virulence immunological studies and vaccination experiments. *Fish Shellfish Immunol.* 2, 211–221.
- Smelty, P., Marchal, C., Renard, R., Sinzelle, L., Pollet, N., Dunon, D., et al. (2010). Identification of the pre-T-cell receptor alpha chain in non-mammalian vertebrates challenges the structure-function of the molecule. *Proc. Natl. Acad. Sci. U.S.A.* 107, 19991–19996.
- Srisapoome, P., Ohira, T., Hirono, I., and Aoki, T. (2004). Genes of the constant regions of functional immunoglobulin heavy chain of Japanese flounder, *Paralichthys olivaceus*. *Immunogenetics* 56, 292–300.
- Star, B., Nederbragt, A. J., Jentoft, S., Grinholz, U., Malmström, M., Gregers, T. F., et al. (2011). The genome sequence of Atlantic cod reveals a unique immune system. *Nature* 477, 207–210.
- Stenvik, J., and Jørgensen, T. Ø. (2000). Immunoglobulin D (IgD) of Atlantic cod has a unique structure. 51, 452–461.
- Stenvik, J., Schröder, M., Olsen, K., Zapata, A., and Jørgensen, T. (2001). Expression of immunoglobulin heavy chain transcripts (VH-families, IgM, and IgD) in head kidney and spleen of the Atlantic cod (*Gadus morhua* L.). *Dev. Comp. Immunol.* 25, 291–302.
- Timmusk, S., Partula, S., and Pilström, L. (2000). Different genomic organization and expression of immunoglobulin light-chain isotypes in the rainbow trout. *Immunogenetics* 51, 905–914.
- Trede, N. S., Langenau, D. M., Traver, D., Look, A. T., and Zon, L. I. (2004). The use of zebrafish to understand immunity. *Immunity* 20, 367–379.
- van Ginkel, F. W., Miller, N. W., Cuchens, M. A., and Clem, L. W. (1994). Activation of channel catfish B cells by membrane immunoglobulin cross-linking. *Dev. Comp. Immunol.* 18, 97–107.
- Vela, J. L., Aït-Azzouzene, D., Duong, B. H., Ota, T., and Nemazee, D. (2008). Rearrangement of mouse immunoglobulin kappa deleting element recombinating sequence promotes immune tolerance and lambda B cell production. *Immunity* 28, 161–170.
- Wang, X., Olp, J. J., and Miller, R. D. (2009). On the genomics of immunoglobulins in the gray, short-tailed opossum *Monodelphis domestica*. *Immunogenetics* 61, 581–596.
- Weill, J.-C., Weller, S., and Reynaud, C.-A. (2009). Human marginal zone B cells. *Annu. Rev. Immunol.* 27, 267–285.
- Weinstein, J. a., Jiang, N., White, R. a., Fisher, D. S., and Quake, S. R. (2009). High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324, 807–810.
- Wienholds, E., Schulte-Merker, S., Walderich, B., and Plasterk, R. H. A. (2002). Target-selected inactivation of the zebrafish *rag1* gene. *Science* 297, 99–102.
- Willett, C. E., Cherry, J. J., and Steiner, L. A. (1997). Characterization and expression of the recombination activating genes (*rag1* and *rag2*) of zebrafish. *Immunogenetics* 45, 394–404.
- Wilson, M., Bengtén, E., Miller, N. W., Clem, L. W., Du Pasquier, L., and Warr, G. W. (1997). A novel chimeric Ig heavy chain from a teleost fish shares similarities to IgD. *Proc. Natl. Acad. Sci. U.S.A.* 94, 4593–4597.
- Wilson, M., Hsu, E., Marcuz, A., Courtet, M., Du Pasquier, L., and Steinberg, C. (1992). What limits affinity maturation of antibodies in *Xenopus*—the rate of somatic mutation or the ability to select mutants? *EMBO J.* 11, 4337–4347.
- Wilson, M. R., van Ravenstein, E., Miller, N. W., Clem, L. W., Middleton, D. L., and Warr, G. W. (1995a). cDNA sequences and organization of IgM heavy chain genes in two holostean fish. *Dev. Comp. Immunol.* 19, 153–164.
- Wilson, M., Ross, D., Miller, N., Clem, L., Middleton, D., and Warr, G. (1995b). Alternate pre-mRNA processing pathways in the production of membrane IgM heavy chains in holostean fish. *Dev. Comp. Immunol.* 19, 165–177.
- Xiao, F. S., Wang, Y. P., Yan, W., Chang, M. X., Yao, W. J., Xu, Q. Q., et al. (2010). Ig heavy chain genes and their locus in grass carp *Ctenopharyngodon idella*. *Fish Shellfish Immunol.* 29, 594–599.
- Yang, F., Waldbieser, G. C., and Lobb, C. J. (2006). The nucleotide targets of somatic mutation and the role of selection in immunoglobulin heavy chains of a teleost fish. *J. Immunol.* 176, 1655.
- Yasuike, M., Boer, J. D., Schalburg, K. R. V., Cooper, G. A., McKinnel, L., Messmer, A., et al. (2010). Evolution of duplicated IgH loci in Atlantic salmon, *Salmo salar*. *BMC Genomics* 11:486. doi:10.1186/1471-2161-11-486
- Ye, J., Kaattari, I. M., and Kaattari, S. L. (2011). The differential dynamics of antibody subpopulation expression during affinity maturation in a teleost. *Fish Shellfish Immunol.* 30, 372–377.
- Zapata, A., Diez, B., Cejalvo, T., Gutiérrez-de Frías, C., and Cortés, A. (2006). Ontogeny of the immune system of fish. *Fish Shellfish Immunol.* 20, 126–136.
- Zhang, Y., Salinas, I., Li, J., Parra, D., Bjork, S., Xu, Z., et al. (2010). IgT, a primitive immunoglobulin class specialized in mucosal immunity. *Nat. Immunol.* 11, 827–835.
- Zhao, Y., Kacskovics, I., Pan, Q., Liberles, D. A., Geli, J., Davis, S. K., et al. (2002). Artiodactyl IgD: the missing link. *J. Immunol.* 169, 4408–4416.
- Zhao, Y., Pan-Hammarström, Q., Kacskovics, I., and Hammarström, L. (2003). The porcine Ig delta gene: unique chimeric splicing of the first constant region domain in its heavy

- chain transcripts. *J. Immunol.* 171, 1312–1318.
- Zhao, Y., Rabbani, H., Shimizu, A., and Hammarström, L. (2000). Mapping of the chicken immunoglobulin heavy-chain constant region gene locus reveals an inverted alpha gene upstream of a condensed epsilon gene. *Immunology* 101, 348–353.
- Zimmerman, A. M., Romanowski, K. E., and Maddox, B. J. (2011). Targeted annotation of immunoglobulin light chain (IgL) genes in zebrafish from BAC clones reveals kappa-like recombining/deleting elements within IgL constant regions. *Fish Shellfish Immunol.* 31, 697–703.
- Zimmerman, A. M., Yeo, G., Howe, K., Maddox, B. J., and Steiner, L. A. (2008). Immunoglobulin light chain (IgL) genes in zebrafish: genomic configurations and inversions rearrangements between (V(L)-J(L)-C(L)) gene clusters. *Dev. Comp. Immunol.* 32, 421–434.
- Zwollo, P. (2011). Dissecting teleost B cell differentiation using transcription factors. *Dev. Comp. Immunol.* 35, 898–905.
- Zwollo, P., Cole, S., Bromage, E., and Kaattari, S. (2005). B cell heterogeneity in the teleost kidney: evidence for a maturation gradient from anterior to posterior kidney. *J. Immunol.* 174, 6608–6616.
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Received: 18 December 2012; paper pending published: 06 January 2013; accepted: 24 January 2013; published online: 13 February 2013.*
- Citation: Fillatreau S, Six A, Magadan S, Castro R, Sunyer JO and Boudinot P (2013) The astonishing diversity of Ig classes and B cell repertoires in teleost fish. *Front. Immun.* 4:28. doi: 10.3389/fimmu.2013.00028*
- This article was submitted to Frontiers in B Cell Biology, a specialty of Frontiers in Immunology.*
- Copyright © 2013 Fillatreau, Six, Magadan, Castro, Sunyer and Boudinot. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.*



The porcine antibody repertoire: variations on the textbook theme

John E. Butler* and Nancy Wertz

Department of Microbiology, Carver College of Medicine, University of Iowa, Iowa City, IA, USA

Edited by:

Harry W. Schroeder, University of Alabama at Birmingham, USA

Reviewed by:

Patrick C. Wilson, University of Chicago, USA

John D. Colgan, University of Iowa, USA

***Correspondence:**

John E. Butler, Department of Microbiology, Carver College of Medicine, University of Iowa, 3-501 BSB, 51 Newton Road, Iowa City, IA 52242, USA.

e-mail: john-butler@uiowa.edu

The genes encoding the heavy and light chains of swine antibodies are organized in the same manner as in other eutherian mammals. There are ~30 VH genes, two functional DH genes and one functional JH gene, 14–60 V κ genes, 5 J κ segments, 12–13 functional V λ genes, and two functional J λ genes. The heavy chain constant regions encode the same repertoire of isotypes common to other eutherian mammals. The piglet models offers advantage over rodent models since the fetal repertoire develops without maternal influences and the precocial nature of their multiple offspring allows the experimenter to control the influences of environmental and maternal factors on repertoire development postnatally. B cell lymphogenesis in swine begins in the fetal yolk sac at 20 days of gestation (DG), moves to the fetal liver at 30 DG and eventually to the bone marrow which dominates until birth (114 DG) and to at least 5 weeks postpartum. There is no evidence that the ileal Peyer's patches are a site of B cell lymphogenesis or are required for B cell maintenance. Unlike rodents and humans, light chain rearrangement begins first in the lambda locus; kappa rearrangements are not seen until late gestation. Dissimilar to lab rodents and more in the direction of the rabbit, swine utilize a small number of VH genes to form >90% of their pre-immune repertoire. Diversification in response to environmental antigen does not alter this pattern and is achieved by somatic hypermutation (SHM) of the same small number of VH genes. The situation for light chains is less well studied, but certain V κ and J κ and V λ and J λ are dominant in transcripts and in contrast to rearranged heavy chains, there is little junctional diversity, less SHM, and mutations are not concentrated in CDR regions. The transcribed and secreted pre-immune antibodies of the fetus include mainly IgM, IgA, and IgG3; this last isotype may provide a type of first responder mucosal immunity. Development of functional adaptive immunity is dependent on bacterial MAMPs or MAMPs provided by viral infections, indicating the importance of innate immunity for development of adaptive immunity. The structural analysis of Ig genes of this species indicate that especially the VH and Cy gene are the result of tandem gene duplication in the context of genomic gene conversion. Since only a few of these duplicated VH genes substantially contribute to the antibody repertoire, polygeny may be a vestige from a time before somatic processes became prominently evolved to generate the antibody repertoire. In swine we believe such duplications within the genome have very limited functional significance and their occurrence is therefore overrated.

Keywords: antibody repertoire, swine, gene duplication, development

THE GENOMIC POTENTIAL FOR THE ANTIBODY REPERTOIRE IN SWINE

The heavy and light chain genome of swine is organized in the familiar translocon fashion of other eutherian mammals, i.e., placental mammals versus egg-layers. The heavy chain locus contains ~30 VH genes, all members of the same VH3 family (Sun et al., 1994). There are five DH segments and five JH segments. However only two DH segments and a single JH are functional (Butler et al., 1996; Eguchi-Ogawa et al., 2010; **Table 1**). Downstream, there are genes encoding C μ , C δ , six subclasses of C γ , C ϵ , and C α . Like cattle, C δ is also associated with a small switch region which may or may not be regularly functional (Zhao et al., 2003; **Figure 1A**). Putative enhancer and promoter elements and switch

regions similar to those described in other mammals have been reported (Sun and Butler, 1997; Eguchi-Ogawa et al., 2010).

The kappa locus is comprised of 14–60 V κ genes in two families, five J κ segments and a single C κ (**Figure 1B**; Butler et al., 2004; Schwartz et al., 2012a; **Table 1**). The lambda locus is comprised of 22 V λ genes, 13 that appear potentially functional and four J λ genes (**Figure 1B**; Schwartz et al., 2012b). Usage of kappa and lambda gene elements is discussed in Section “The Pre-Immune Antibody Repertoire of Swine.” The organization of these loci is generally conserved among eutherian mammals although the kappa locus is duplicated in rabbits (**Table 1**). It is not the purpose of this article to provide a phylogenetic review of the Ig loci of mammals and other vertebrates. We mention only selected

Table 1 | Diversity in genomic potential for antibody diversity in common mammals.

Species	V _H (F ^a)	D _H	J _H	V _λ (F)	J _λ	C _λ ^b	V _κ (F)	J _κ	C _κ	κ:λ ^c
Human	87 (7)	30	9	70 (7)	7	7	66 (7)	5	1	60:40
Mouse	>100 (14)	11	4	3 (3)	4	4	140 (4)	4	1	95:5
Rabbit	>100 (1)	11	6	? (?)	2	2	>36 (?)	8	2 ^d	95:5
Horse	>10 (2)	>7	>5	25 (3)	4	4	>20 (?)	5	1	5:95
Cattle	>15 (2)	3	5	83 (8)	>2	4	? (?)	?	1	5:95
Swine	>20 (1)	2 ^e	1 ^e	22(>2)	>4	4	14–60 (5)	5	1	50:50
Bat	>250 (>5)	?	13	? (?)	?	?	? (?)	?	?	??

^aNumber of families (F) of variable region genes.

^bJ_λ-C_λ dupicons are the common motif in most mammals.

^cRatio of expressed light chain in adults expressed as percent.

^dRabbits have a duplicate of the entire kappa locus.

^eFunctional D_H and J_H genes.

?, Number is unknown.

species as a reference to help readers less familiar with comparative immunology.

THE PIGLET MODEL FOR STUDIES OF ANTIBODY REPERTOIRE DEVELOPMENT

The piglet provides an ideal model for studies on antibody repertoire development for a number of important reasons. First, swine are members of the hooved mammal group, i.e., Ungulates, which have a form of placentation that, unlike that in rodents, primates, and rabbits, does not allow transport of maternal antibodies and other proteins *in utero* to the developing fetus (Brambell, 1970; Butler, 1974). Gestation is 114 days which allows 84 days from the time that VDJ rearrangements first appear to study the development of B cells and the antibody repertoire during fetal life in their multiple large fetuses. Because of the placentation described, development during this period is considered intrinsic and not regulated by maternal factors transmitted *in utero*. Second, the offspring of swine and all Ungulates are Precocial. In the context of development, this refers to the ability of newborn Ungulates to be born totally mobile, with fully functional eyes and protective fur/hair and the ability to immediately forage. Thus piglets can be recovered by Caesarian surgery and reared separately from their mothers in germfree isolators or SPF autosows (Butler et al., 2009a). This allows the experimenter to control the exposure of the postnatal offspring to maternal factors, commensal flora, certain dietary regimes, and infectious agents.

The swine has another major advantage; seven major VH genes, all of the same family, two DH segments and a single JH account for >90% of the VDJ repertoire (Butler et al., 1996; Sun et al., 1998). This is true in the yolk sac (YS) at 20 days of gestation (DG) and continues into adulthood (Butler et al., 2006a, 2011a). Thus all VDJ rearrangements can be recovered from DNA or from transcripts using a single PCR primer set. Cloning these rearrangements and using probes that recognize the CDR regions of each major gene as well as sequence analysis allows vertical studies on the developing repertoire and for its quantification according to a repertoire diversification index (RDI; Butler et al., 2006a, 2011a).

These features collectively allow factors that act during the “critical window” of immune development to be addressed (Figure 2).

The critical window in the swine system is the period when innate immunity, “natural antibodies” and passive immunity gain the help of the developing adaptive immune response system to allow offspring survival. Survival of the newborn through this critical period when adaptive responses are poorly developed depends on passive immunity in which the systemic humoral experience of the mother is transmitted via IgG and the mucosal experience by IgA. This is also the time when neonatal tolerance to non-threatening dietary antigens and commensal gut flora become established. These topics are discussed in detail in other reviews (Butler, 1983; Butler and Kehrle, 2005). Relevant to the theme of this chapter is the role played by colonizing gut flora on the development of adaptive immunity and on the diversification of the antibody repertoire.

B CELL DEVELOPMENT IN SWINE

Any discussion of antibody repertoire development requires some mention of the B cell lineage from which antibodies are derived. In the fetus, VDJ rearrangements are first seen in YS at 20 DG (Sinkora et al., 2003; Sun et al., 2012a). However signal joint circles (SJC) are difficult to recover at this time, probably because of the slow rate of B cell lymphogenesis at this early time and the rapid rate of their degradation (Figure 3). TdT is expressed and active at this time but N region additions and CDR3 diversity is low compared to older fetuses and young piglets (Sinkora et al., 2003; Butler et al., 2007; Sun et al., 2012a). B cell lymphogenesis moves to the fetal liver at 30 DG and remains active in this organ until the bone marrow (BM) develops at ~65 DG. SJC are readily detectable at all these sites indicating that it unlikely the B cells found in these sites are immigrants. An interesting feature of heavy chain rearrangements in early B cell lymphogenesis is the unexpected high frequency (>85%) of in-frame rearrangement (Sinkora et al., 2003), suggesting that the rearrangement and selection processes differs between early fetal stages of B cell lymphogenesis and those which occurs later in fetal and adult BM. However, evidence for B-1 and B-2 subpopulations as reported in mice (Herzenberg et al., 1986) has so far not been obtained. We also found no evidence that the ileal Peyer patches (IPP) are a site of B cell lymphogenesis or that they are needed for B cell

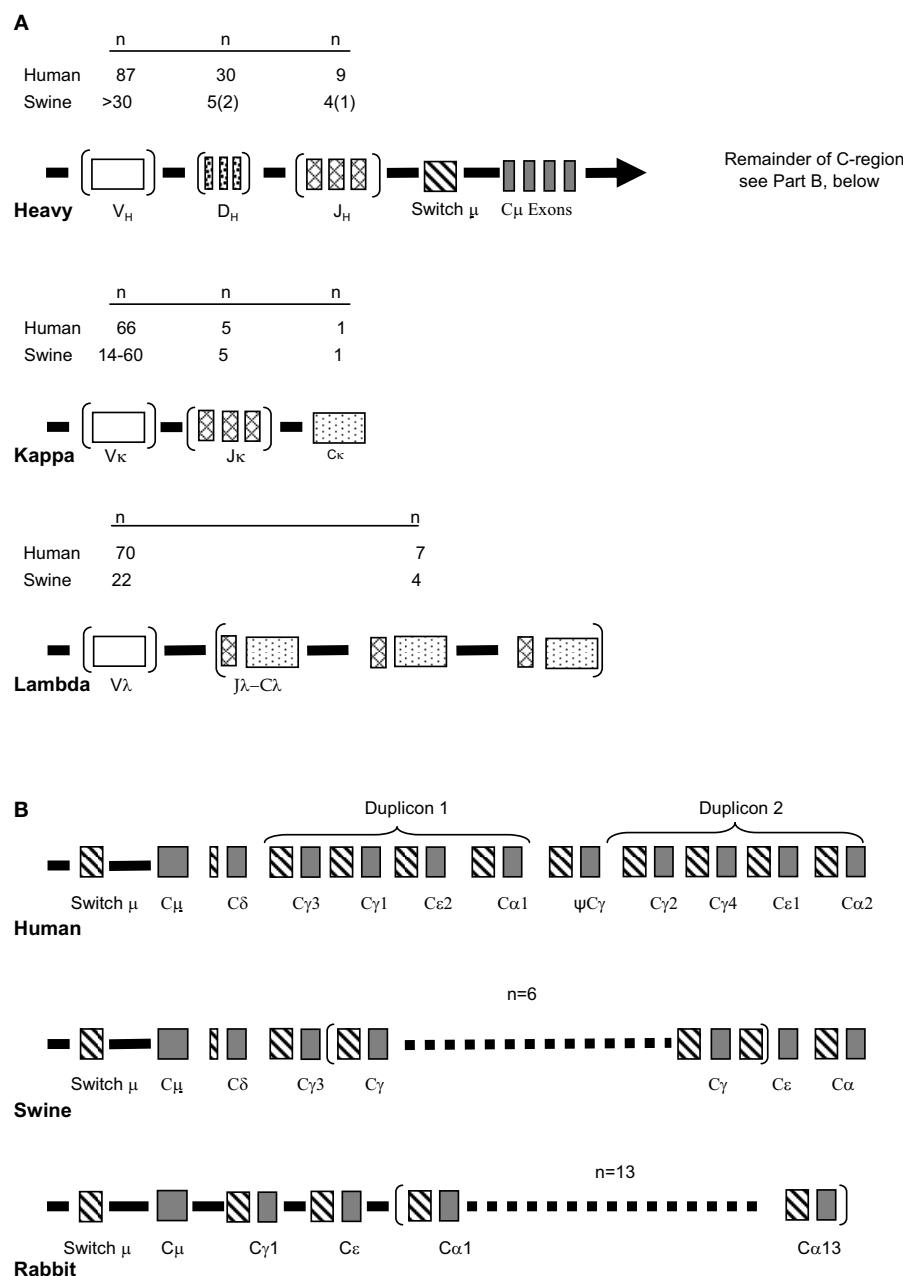


FIGURE 1 | The genomic repertoire of heavy and light chain genes of swine. (A) The light chain and heavy chain variable gene loci compared to humans. Dupicons are illustrated in parentheses with the actual number given in the mini-table above designated as “n.” This includes dupicons of

cassettes as in the lambda locus. In the swine only two DH and one JH gene cassettes are functional (given in parenthesis). **(B)** Variation in the organization of the heavy chain constant region among human, swine, and rabbit. Modified from Butler et al. (2011c).

maintenance (Butler et al., 2011b; Sinkora et al., 2011; Sun et al., 2012a).

A major departure from the pattern seen in lab rodents and humans is the order of light chain rearrangement. Rearrangement begins in the lambda locus at 20 DG in YS and is dominant until in late gestation when kappa rearrangements first appear (Figure 3; Sun et al., 2012a). This may not be unusual given that in Ungulates, lambda usage dominates the repertoire; in cattle, sheep, and

horse and it can exceed >90% (Butler, 1997; Table 1). However usage of light chain isotypes in young pigs and adults is roughly equal, similar to humans (Hood et al., 1967; Skvaril et al., 1976; Butler et al., 2005a; Sun et al., 2012a; Table 1). The order of light chain gene segment usage during development has been poorly studied in any Ungulate. However, the IGLV8 ($V_\lambda 8$) family is exclusively used at early sites of B cell lymphogenesis (Vazquez et al., 2012).

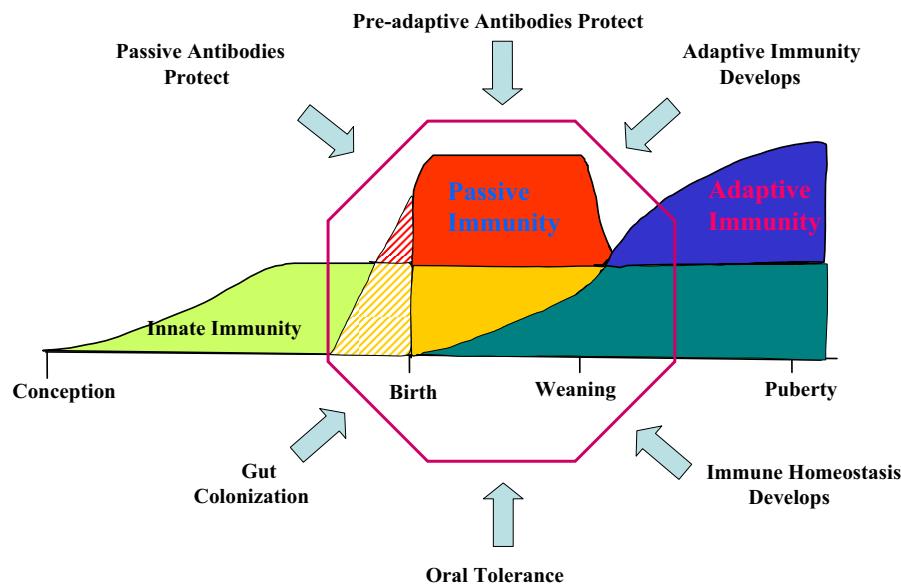


FIGURE 2 | The critical window of immunological development. The factor affecting events in the “window” are indicated. There are a number of cases of superimposing events; e.g., Passive immunity (red) superimposed on innate immunity (green) produces yellow. The “lined

section” prepartum applies only to species, e.g., mice and humans, in which passive immunity also takes place *in utero*. In Ungulates like swine there is no transfer of passive antibodies *in utero*. Modified from Butler et al. (2006b).

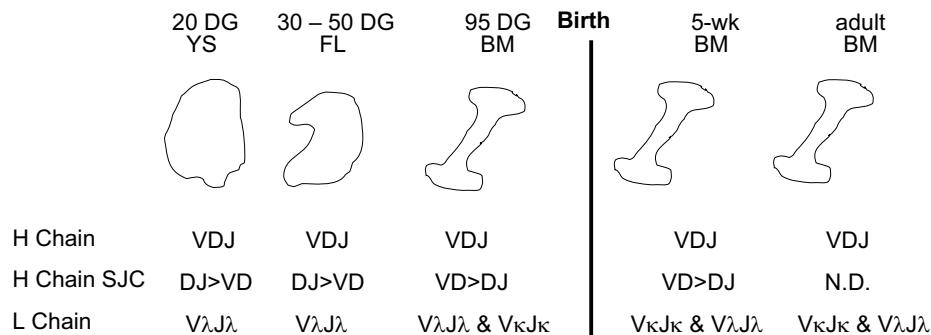


FIGURE 3 | B cell lymphogenesis in swine. Heavy and light chain rearrangements and signal joint circles (SJC) recovered from. YS, yolk sac; FL, fetal liver, BM, bone marrow; DG, day of gestation; N.D., not detected.

It has been proposed that B cell development among higher vertebrates places species into groups (Lanning et al., 2004). In rodents and primates the BM is the major site of B cell development and diversification whereas in rabbit and sheep, gut associated lymphoid tissues (GALT) are believed to be critical (Lanning et al., 2004; Mage et al., 2006). However, recent findings indicate that swine do not belong to the GALT group (Butler et al., 2011b; Sinkora et al., 2011). This is based on surgical resection of IPP, which did not affect B cell lymphogenesis, repertoire diversification and maintenance of either B or T cell levels. The swine IPP appears to be merely a type of mucosal immune tissue.

Some level of class-switch recombination (CSR) occurs during fetal life, so that IgM, IgG3, and IgA are transcribed and secreted into serum. However, no environmental antigen is present during this time and it is unknown whether this CSR involves germinal

center formation. There is also low level somatic hypermutation (SHM) in rearranged VDJs resulting in <10 mutation per kilobase (Butler et al., 2011a). In heavy chain rearrangement these mutations accumulate in CDR regions. In lambda rearrangement mutations are equally distributed in CDR and FR regions, although the total mutation frequency is 10-fold lower than in heavy chain rearrangements (Vazquez et al., 2012). Tests for expression of AID during fetal life have not been undertaken.

THE PRE-IMMUNE ANTIBODY REPOertoire OF SWINE

We define the pre-immune repertoire as the one present in fetal and newborn piglets. Conventionally reared swine have been previously exposed to both environmental antigen and regulatory influences from maternal passive antibodies, but isolator piglets are free of such influences. In their circulation, newborns have ~30 µg/ml

of IgG, 1 µg/ml of IgM, and 2 µg/ml of IgA (Butler et al., 2001, 2009b). If maintained for 5 weeks in germfree conditions in which dietary protein is the sole source of environmental antigen, IgM and IgA levels increase 3- to 5-fold to ~6 µg/ml, but there is only a 20% increase in IgG to 40–45 µg/ml. Since these animals are immuno-unresponsive (see below) we suspect this small increase is an intrinsic developmental effect much as is the CSR and SHM that occur during fetal life. We base this on the fact that isolator piglets that become immunoresponsive have a 20- to 35-fold increase in serum Ig across all major isotypes. Of the six subclasses of IgG, IgG3 accounts for >60% of the C_γ transcripts in mucosal tissues during the period prior to environmental exposure (Butler and Wertz, 2006).

A notable variation on the theme of repertoire development and diversification from what is presented in textbooks and reviews concerns VH gene usage. Swine use 7 VH genes to form nearly their entire VDJ repertoire starting first in YS and continuing throughout gestation and beyond into postnatal life (Figure 4; Butler et al., 2011a). With minor exceptions, proportional usage remains constant and the usage does not depend on the position of the VH gene in the genome (Eguchi-Ogawa et al., 2010; Butler et al., 2011a). VHG (IGHV2) which is the most 3' functional gene is seldom used whereas VHN (IGHV15) can account for 13% of the repertoire in YS at 20 DG (Butler et al., 2011a; see Antibody Repertoire Development and the Origin of the Genomic Repertoire). Exceptions to the constancy of VH usage involves decreased usage of VHN in older fetuses and a reciprocal increase in usage of VHC. As will be discussed in more detail in the next section (see Development of Adaptive Immunity Depends on an Encounter with MAMPs) this constancy of VH usage continues after birth in antigenized animals including those reared conventionally and consequently exposed to a plethora of environmental antigens. There is no consistent change in the frequency of usage between the two DH segments and swine have only one functional JH (Butler et al., 1996; Eguchi-Ogawa et al., 2010). The small number of VH, DH, and JH segments used means that combinatorial diversity is very small, i.e., 14 possibilities compared to ~9000 in humans.

Thus, we estimated that junctional diversity in CDR3 accounts for >95% of the swine pre-immune repertoire (Butler et al., 2000a). As indicated above, the frequency of SHM remains low during this period but the complexity of CDR3 is high and the Gaussian spectratype pattern of CDR3 lengths suggests an unselected repertoire (Navarro et al., 2000; Butler et al., 2007).

As indicated in Section “The Genomic Potential for the Antibody Repertoire in Swine,” the overall genomic structure of the kappa and lambda loci of swine is similar to that in other well-studied mammals. Genomic gene annotation may not totally predict the expressed repertoire. In the case of kappa, 11 functional V_κ genes and 5 J_κ genes are present in the genome, yet 2 V_κ families and 1 J_κ segment account for most of the repertoire (Butler et al., 2004; Schwartz et al., 2012a). In the case of lambda, V_λ usage appears to closely agree with the genomic potential but only two of the four J_λ genes are used (Schwartz et al., 2012b; Vazquez et al., 2012).

Certain features of the pre-immune light chain repertoire differ substantially from those in the heavy chain but these features are not unique to swine. First, SHM is 10-fold lower in light chain rearrangement than in heavy chain rearrangement of piglets from the same population. Furthermore there is little junctional diversity in kappa or lambda rearrangement and CDR3 length is ~30 ± 2 (Butler et al., 2004; Vazquez et al., 2012). These observations are nearly identical to studies in humans (Victor et al., 1994; Bridges et al., 1995; Girschick and Lipsky, 2001; Richl et al., 2008). Of some notice is that in the pre-immune repertoire, SHM is concentrated in the CDR region of heavy chain rearrangements but both are lower and widely distributed in rearranged light chains (Butler et al., 2006b, 2011a; Vazquez et al., 2012). The significance of this difference is unclear. However, one might speculate that because the antibody binding site is primarily determined by the heavy chain and its CDR3 (Padlan, 1994) light chain may provide only a supporting role and their presence primarily affects the conformation of the heavy chain binding site. Their complete absence in the camelids partially supports this view (Nguyen et al., 2002).

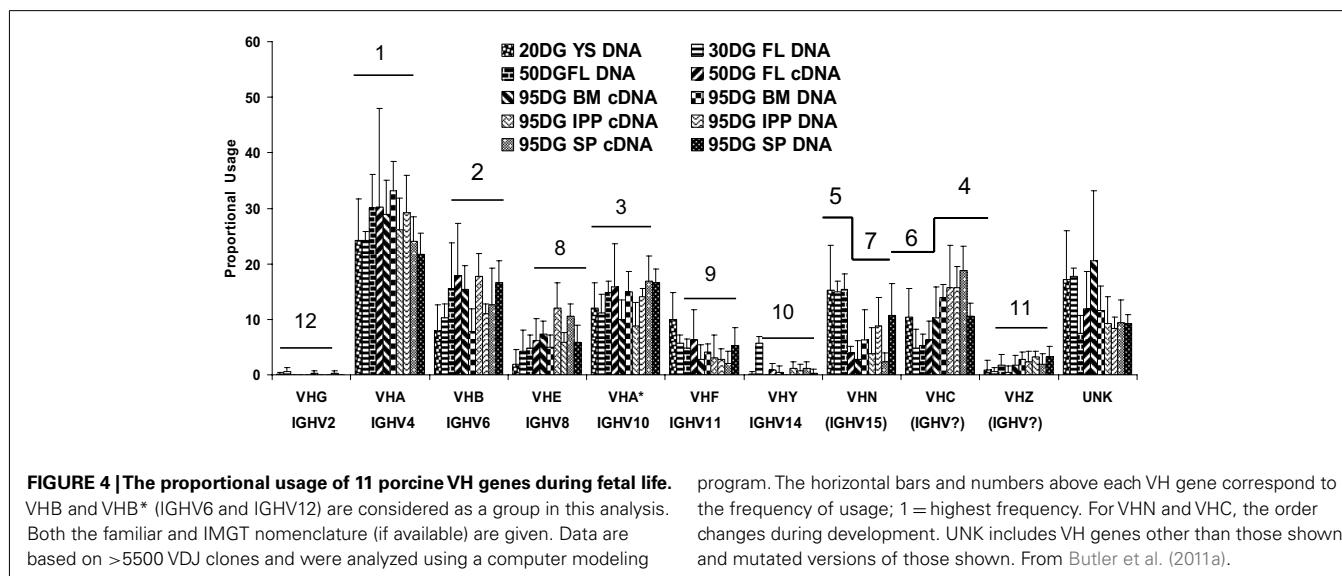


FIGURE 4 |The proportional usage of 11 porcine VH genes during fetal life. VHB and VHB* (IGHV6 and IGHV12) are considered as a group in this analysis. Both the familiar and IMGT nomenclature (if available) are given. Data are based on >5500 VDJ clones and were analyzed using a computer modeling

program. The horizontal bars and numbers above each VH gene correspond to the frequency of usage; 1 = highest frequency. For VHN and VHC, the order changes during development. UNK includes VH genes other than those shown and mutated versions of those shown. From Butler et al. (2011a).

DEVELOPMENT OF ADAPTIVE IMMUNITY DEPENDS ON AN ENCOUNTER WITH MAMPs

COLONIZATION AND MAMPs

Piglets maintained germfree in isolators for 5–6 weeks show only minor changes in serum Ig levels until after colonization (Butler et al., 2000b, 2009b; see The Pre-Immune Antibody Repertoire of Swine), do not have antibodies to dietary proteins (J.E. Butler and Patrick Weber, unpublished observations) and are unable to respond to T cell dependent (TD) and T cell independent (TI-2) antigens. However, colonization with benign *Escherichia coli* or a probiotic cocktail, allows responses to both types of antigens (Butler et al., 2002). In lieu of living bacteria, purified MAMPs (bacterial DNA as CpG-ODN, muramyl dipeptide or LPS) have the same effect (Butler et al., 2005b). Thus, bacterial MAMPs provide the adjuvant necessary for innate immune receptors to stimulate the development of adaptive immunity. The impact of such exposure results in 100- to 1000-fold increase in serum IgGs (Butler et al., 2009b), CSR to downstream Cy genes, (Butler et al., 2012a) a 3- to 5-fold increase in the frequency of SHM and a 1–2 log increase in the RDI (Butler et al., 2011a). CpG-ODN and LPS are polyclonal B cell activators and can also expand the existing B cell populations to secrete IgM, IgA, and IgG3 antibodies. However, such expansion cannot be considered a “somatically adapted” repertoire.

REPERTOIRE DIVERSIFICATION FOLLOWING INFECTION WITH RNA VIRUSES

Viruses have a broad range of effects on adaptive immunity. Some are polyclonal activators while others suppress immune responses by interfering with antigen presentation by a variety

of mechanisms (Coutelier et al., 1990; Ehrlich, 1995; Hahn et al., 1998; Acha-Orbea et al., 1999; Hunziker et al., 2003). However some, such as influenza (FLU), stimulate robust antibody responses, the apparent basis of generally high efficacy FLU vaccines. Such viruses generate dsRNA during replication, a known adjuvant (Cunnington and Naysmith, 1975). In piglets, we have studied three pandemic viruses including swine influenza (S-FLU) and another RNA virus called porcine respiratory and reproductive syndrome virus (PRRSV) which acts as a polyclonal activator of B cells in both germfree and colonized piglets and fetuses inoculated *in utero*. Polyclonal activation by PRRSV results in lymphoid adenopathy, hypergammaglobulinemia, the appearance of autoantibodies and the deposition of immune complexes in kidney (Lemke et al., 2004). Infection with PRRSV expands certain B cells clones that display hydrophobic CDR3s, a feature common to antibodies that comprise the pre-immune repertoire (Butler et al., 2007, 2008; Schelonka et al., 2007). In addition to polyclonal activation there is also diversification of the repertoire, not dissimilar from that seen in piglets infected with S-FLU or colonized with gut flora (**Figure 5**). However, the degree of repertoire diversification does not parallel the increase in serum IgGs (Sun et al., 2012b). This results in only a very small proportion of virus-specific IgGs (Lemke et al., 2004).

Infection of isolator piglets with S-FLU results in a robust IgG response to the virus but a much weaker response to irrelevant model TI-2 and TD antigen than does gut colonization (Butler et al., 2012a). Thus, S-FLU does not have the same robust adjuvant impact as does bacterial colonization. This may be due to the fact that S-FLU offers primarily one TLR ligand, double-stranded

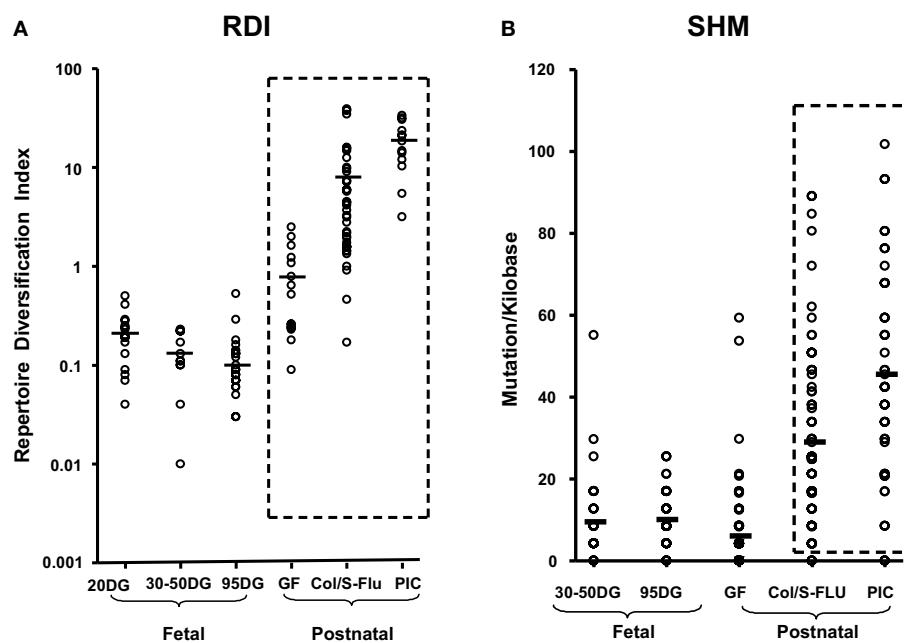


FIGURE 5 | Diversification of the porcine antibody repertoire during fetal and postnatal life expressed as a repertoire diversification index [RDI; (A)] or as the frequency of somatic hypermutation [SHM; (B)]. DG, days of gestation. Values for 95 DG are pooled from clone frequencies recovered

from spleen, IPP, and MLN since there were no tissue differences. GF, germfree isolator piglets; Col S-FLU, colonized or S-FLU infected isolator piglets; PIC, parasite-infected young adults reared conventionally. From Butler et al. (2011a).

RNA or that the gut innate immune system is more responsive than that of the respiratory tract. In any case, the result of S-FLU infection supports the concept that MAMPs awakens the adaptive immune system.

A third pandemic disease of swine is porcine circovirus Type 2, a small DNA virus (PCV2; Allan and Ellis, 2000; Merial, 2004). In the piglet model, PCV2 has little or no adjuvant effect in terms of stimulating the production of antibodies to irrelevant model TI-2 and TD antigens. However infection with PCV2 nevertheless results in a generally robust response to a recombinant ORF2 antigen of the virus (Sun et al., 2012c). This IgG response in serum occurs in the context of elevated IgA levels in serum and bronchial-alveolar lavage (BAL). PCV2 infection does not result in polyclonal B cell activation, rather it targets IgA-producing cells. These findings further emphasize that the humoral immune response to any one particular viral infection, should not be projected to other viral infections.

DIVERSIFICATION OF THE HEAVY VARIABLE REGION REPERTOIRE IS BY SHM OF VH GENES THAT COMPRIZE THE PRE-IMMUNE REPERTOIRE

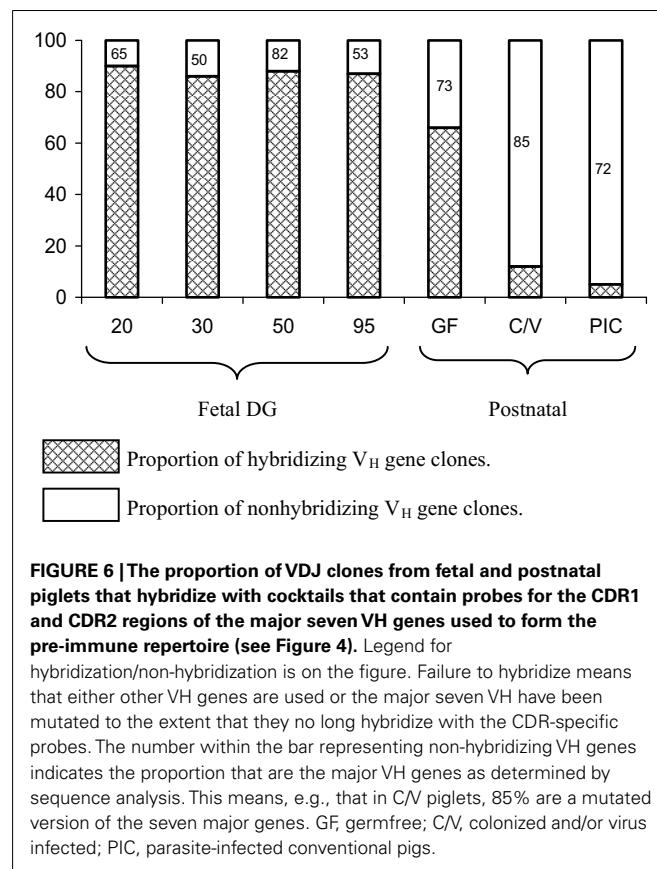
The user friendly and simple VDJ system of swine combined with the other advantages of the piglet model (see The Piglet Model for Studies of Antibody Repertoire Development) make it possible to follow the developmental history of the seven major VH genes that comprise >90% of the pre-immune repertoire in postnatal animals under various environmental conditions. The surprising result is that exposure to environmental antigen does not result in the recruitment of other VH genes from the genomic repertoire that are not present in the pre-immune repertoire. Rather, somatic mutants of the “magnificent seven” VH genes that comprise the pre-immune repertoire comprise the adaptive repertoire (**Figure 6**). In piglets infected with S-FLU, colonized with bacteria or helminth parasites, the same VH genes comprise the repertoire but ~90% of them are somatically mutated (Butler et al., 2011a; **Figure 6**). The use of somatic gene conversion, prominent in the rabbit (Knight, 1992; Schiaffella et al., 1999; Winstead et al., 1999) has not been observed in swine.

ENVIRONMENTAL EXPOSURE RESULTS IN CSR TO DOWNSTREAM C_γ GENES

We described above that in naïve newborns and fetal piglets, >60% of IgG transcripts encode IgG3. However, fetal infection or postnatal exposure to virus, normal gut flora or parasitic infection reduces IgG3 transcription to ~5% (Butler et al., 2012a). The resultant IgG is now encoded by downstream C_γ genes of which IgG1 is a major player (Butler and Wertz, 2006). The switch to downstream C_γ genes (**Figure 7A**) parallels the diversification of their VH genes while the repertoire associated with IgM and IgG3 does not diversify (**Figure 7B**). These observations are further evidence that environmental exposure of fetal and newborn piglets turns on the machinery of the adaptive immune system. However, since IgA and IgG3 are already transcribed and secreted by the fetus, environmental exposure is not obligatory for CSR.

THE SWINE VERSUS ESTABLISHED PARADIGMS OF ANTIBODY REPERTOIRE DEVELOPMENT

The paradigms for development of the mammalian antibody repertoire decorate the pages of immunology textbooks and are



embodied in a number of classic reviews (Rajewsky et al., 1987; Cohn and Langman, 1990). Discussed in this section is how well the swine system fits these paradigms. In the swine system, some level of CSR and SHM occurs in the absence of environmental exposure. This may also be the case in lab rodents and in human, but is ambiguous in these species because such changes could result from the regulatory effects of maternal antibodies (Wikler et al., 1980; Rodkey and Adler, 1983; Yamaguchi et al., 1983; Wang and Shlomchik, 1998) or may be the result of antigen trafficking across the maternal-fetal barrier (Tristram, 2005). In any case, newborn piglets enter the world with a “natural antibody repertoire” of IgM, IgA, and IgG3 antibodies; a phenomenon that is probably universal among all vertebrates (Ochsenbein and Zinkernagel, 2000).

In swine, the evolution of antibody repertoire development appears to have followed a somewhat different path than in mice and humans. However, we have been unable to identify vertical studies on VH usage in rodents or humans, equivalent to those done in swine to thoroughly confirm these differences (Sun et al., 1998; Butler et al., 2011a). In swine, proportional VH gene usage appears to be constant from the time of the initial B cell development in the YS to adulthood (Butler et al., 2011a). Adaptive diversification of the repertoire depends on SHM of the same major VH genes that comprise the pre-immune repertoire. In humans and mice, adaptive responses are often ascribed to the selective use of certain VH genes (Sheehan et al., 1993; Glas et al., 2000). Despite these differences between mice and swine,

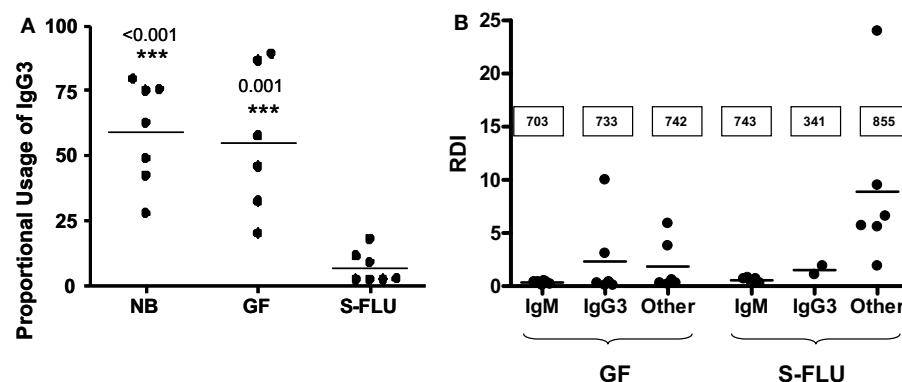


FIGURE 7 | Adaptive changes to the antibody repertoire as a result of postnatal infection with S-FLU. (A) Effect on IgG3 transcription in the tracheal-bronchial lymph node (TBLN). NB, newborn; GF, 5-week germfree isolator piglets; S-FLU, 5-week S-FLU infected isolator piglets. (B) The

repertoire diversification index (RDI) for VH genes transcribed with IgM, IgG3 and downstream Cy transcripts (called: “other”) in GF and S-FLU infected piglets. The boxed values are the number of clones examined. Data indicate that the IgM and IgG3 repertoire does not diversify in S-FLU infection.

there are many similarities. For example, repertoire diversification is by SHM with mutations accumulating in the CDR regions (Berek and Milstein, 1987; Butler et al., 2006a, 2011a) with no evidence of somatic gene conversion as has been reported in the rabbit and chicken (Reynaud et al., 1987; Becker and Knight, 1990; Schiaffella et al., 1999; Winstead et al., 1999; Ratcliffe, 2006). Consistent with studies in humans, light chains provide limited diversity (see The Pre-Immune Antibody Repertoire of Swine). Also similar to mice and human is the duplication of the Cy genes and the changes in their expression upon antigenic stimulation (Mossman and Coffman, 1989). This of course differs from lagomorphs that have a single Cy gene but 13 genes for IgA that comprise their repertoire (Burnett et al., 1989; Figure 1). Like humans and rodents, there is no apparent equivalent to the specialized IgG1 of ruminant artiodactyls that is believed to be essential for passive immunity from mother to young, and which also appears to function as a mucosal antibody (Butler, 1983; Butler and Kehrle, 2005). In regard to mucosal immunity, swine have essentially the same IgA-dependent system as rodents and humans including a well-developed gut-mammary gland axis. Both swine and rodents lack the specialized long-hinged IgA1 of humans and primates, which is especially susceptible to bacterial proteases (Plaut et al., 1974) although *H. suis* produces an unrelated protease that cleaves both porcine IgA allotypic variants and may well cleave the IgA of most mammals (Mullens et al., 2011).

At this point, information on diversification of the light chain repertoire of swine may be inadequate. At this time there is little to suggest that only a small number of V λ or V κ genes comprise the diversified repertoire of the light chains in the manner we have described for the VH genes in this species. Rather, a much larger array of V κ and V λ genes are used (Butler et al., 2004; Vazquez et al., 2012). However, like mice and humans, length junctional diversity in the light chain repertoire is restricted (Victor et al., 1994; Bridges et al., 1995; Girschick and Lipsky, 2001; Richl et al., 2008). But unlike the camelids, another Ungulate, light chains have not become obsolete (Hammers-Casterman et al., 1993; Nguyen et al., 2002). There is much to favor the idea that light chains help

stabilize the heavy chain binding site and allow specificity modification and therefore the rescue of autoreactive B cells through receptor editing (Tiegs et al., 1993).

As regards the adjuvant effect of bacterial and viral MAMPs that act on innate immune receptors, this is most likely a universal phenomenon among higher vertebrates that go on to develop an effective adaptive immune system. In any case, this topic falls outside the main theme of this review.

ANTIBODY REPERTOIRE DEVELOPMENT AND THE ORIGIN OF THE GENOMIC REPERTOIRE

ORIGIN AND IMPORTANCE OF THE GENOMIC REPERTOIRE

Discussion of the expressed antibody repertoire should also consider its genomic origin. The presence of an annotated gene is not alone evidence for its function; the many breeds of dogs are a poignant example. Insight into why the genomic repertoire of Ig genes greatly exceeds the functional repertoire may help to explain their phylogeny. This can help to explain the redundancy that is a feature of the genomic repertoire which can assure an effective adaptive immune system among higher vertebrates.

DUPLICATION AND GENOMIC GENE CONVERSION EXPLAINS THE PORCINE VH AND Cy GENOMIC REPERTOIRE

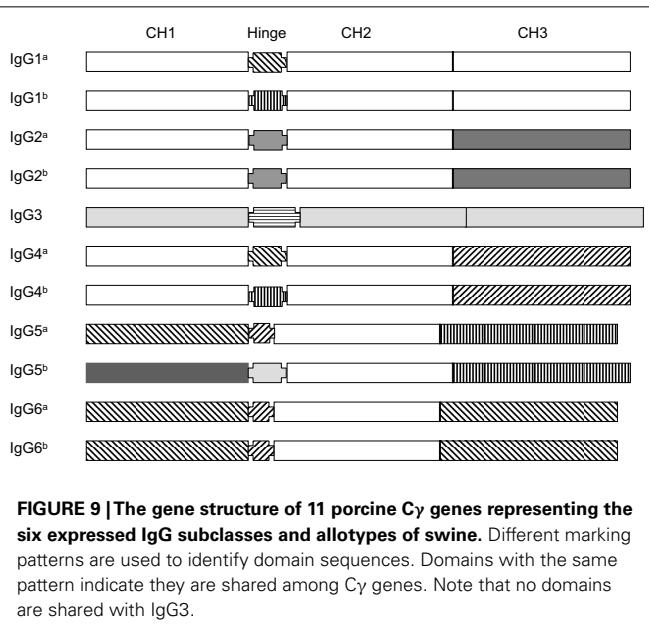
Figure 8 aligns the sequences of a number of porcine VH genes to show that with minor exceptions, they only differ in their CDR1 and CDR2 regions. Thus, the genomic repertoire is a “mix and match” CDR potpourri, with CDR regions shared among different VH genes. Certain VH genes like VHA and VHA* (IGHV4 and IGHV10) and VHB and VHB* (IGHV6 and IGHV12) are duplications with several mutations; one in the CDR1 region of VHB* and two in FR3 of VHA*. In fact there is evidence that the genes segments encoding these duplicated VH genes and several others were duplicated as a block (Eguchi-Ogawa et al., 2010) similar to the dupicons in the human heavy chain constant region and the J λ C λ loci of all studied mammals (Figure 1B). These features of duplicated genes are also seen among the VH genes of bats (Bratsch et al., 2011; Butler et al., 2011d). These comparisons support the hypothesis that the germline VH repertoire in swine

Vernacular	IGMT	FR1	CDR1	FR2	CDR2	FR3
$V_{H}G$	IGHV2		G		E	
$V_{H}A$	IGHV4		A		A	I
$V_{H}T$	IGHV5		E	I	F	I
$V_{H}B$	IGHV6		I	B	B	II
$V_{H}E$	IGHV8		E		E	
$V_{H}A^*$	IGHV10		A		A	
$V_{H}F$	IGHV11		I	F	I	
$V_{H}B^*$	IGHV12		I	P	B	I
$V_{H}Y$	IGHV14		C		A	II
$V_{H}N$	IGHV15		N		N	II
$V_{H}L$	Not mapped		I	E	L	II
$V_{H}I$	Not mapped		E		I	I
$V_{H}J$	Not mapped		J		F	I
$V_{H}Z$	Not mapped		I	E	C	I
$V_{H}C$	Not mapped		C		C	I

FIGURE 8 | Evidence in support of duplication of VH genes in swine.

Diagrammatic comparison of 16 of 32 VH3 family genes described for swine. The sequences are drawn to approximate scale. CDR sequences are identified by capital letter. Only the first fifteen 3' VH genes have been

mapped; pseudogenes are not shown. Those VH genes with the same A, B, etc. CDR designations share the same CDR1 or CDR2 regions. In the case where small differences (genomic point mutations) in FR and CDR regions are involved, vertical lines are drawn.

**FIGURE 9 | The gene structure of 11 porcine Cy genes representing the six expressed IgG subclasses and allotypes of swine.** Different marking patterns are used to identify domain sequences. Domains with the same pattern indicate they are shared among Cy genes. Note that no domains are shared with IgG3.

results from gene duplication that occurred simultaneously with genomic gene conversion. The same mechanism appears evident when the sequences of the six expressed porcine Cy genes and their allotypic variants are compared (Figure 9). For example, the allotypic variants of IgG1 and IgG4 have the same hinge exons and the hinge of IgG5^a is shared with the allelic variants of IgG6. With exception of IgG5^b and IgG3, the Cy1 domain of all Cy subclass genes and their alleles are identical.

Dendrogram analyses of sequence data that are often called phylogenetic studies, indicate that IgG3 is the ancestral IgG for swine. This and other evidence indicates that subclass diversification

occurred after speciation (Kehoe and Capra, 1974; Nguyen, 2001; Butler et al., 2009c). We believe the other five Cy genes then diversified from IgG3. Convincing evidence could be best obtained by studying the Cy subclasses of other swine related species that are related to the ancestors of the domesticated pig. Consistent with studies on human Cy (LeFranc et al., 1991), there is also evidence for deletion of some Cy genes in some pigs (Butler and Wertz, 2006; Eguchi-Ogawa et al., 2012).

THE RESTRICTED USE OF VH GENES IN SWINE QUESTIONS THE IMPORTANCE OF COMBINATORIAL DIVERSITY

As shown in Table 1, lab rodents, rabbits, humans, and bats have and/or express large numbers of VH genes; the same is true for the more primitive Zebrafish (Weinstein et al., 2009). In swine and ruminant artiodactyls, the number is surprisingly low. In swine with ~30 VH genes, only seven appear to be used/needed for a healthy adaptive immune system. Many of the others may represent only allelic variants, so the actual number of germline VH genes may be <20. This raises a question about the textbook paradigms of the importance of multiple VH genes and combinatorial diversity in generating a protective antibody repertoire. This was tested by Xu and Davis (2000) who used a mouse with a single functional VH gene, but with an intact DH and JH region. They showed this transgenic mouse could make antibodies to nearly all environmental antigens. This can be interpreted to mean that extensive duplication of VH genes is not required for the ability to make antibodies to many specificities. This may explain why ruminant artiodactyls and swine have a small genomic repertoire yet are among the most successful mammals on the planet (Table 1). However the Xu and Davis work also suggests that the critical feature is DH and JH polygeny, which invokes the next question of how swine generate diversity with only two functional diversity segments and a single JH. This brings the focus to junctional

diversity in the formation of CDR3 and away from the emphasis on the number of functional VH, DH, and JH segments. CDR3 is considered most important for the specificity of antibodies (Padlan, 1994), whether generated from a genome with many available DH and JH segments, or from a genome with just one or several DH and JH segments. Perhaps this is compensated by light chain diversity, yet the camelids do it without the assistance of light chains (Hamers-Casterman et al., 1993; Nguyen et al., 2002).

IS THE DIVERSIFICATION OF THE IgG SUBCLASSES IN MAMMALS REALLY NECESSARY FOR SPECIES SURVIVAL?

Textbooks and reviews emphasize the importance of the division of labor among different antibody isotypes; each constant region permitting some special biological function, i.e., following the accepted concept of structure function relationship. While this is clear for IgM, IgE, IgA, and IgG, applying this to the subclasses of IgG is less convincing, especially in species like the horse, swine, and bats that have large numbers of Cy variants while the highly successful rabbit lacks subclass variants. We also know that IgD knockout mice behave normally (Nitschke et al., 1993), rabbits lack IgD altogether (Lanning et al., 2003) and even mammals with a gene for IgD apparently do not express it. In humans, deficiencies of individual IgG subclasses are well known but these have not translated to an effect on human health, even in environmentally stress underdeveloped countries (LeFranc et al., 1983; Olsson et al., 1993; Rabbani et al., 1995). In mice, differential IgG subclass expression is related to the balance between inflammatory and regulatory cytokine (Mossman and Coffman, 1989) but there is little evidence that the resulting subclasses make a difference in protective immunity; i.e., does the absence either of IgG1 or IgG2b antibodies really affect protective immunity? Four decades of phenomenological studies in the veterinary world would suggest that the IgG1 subclass is indispensable for passive immunity in ruminant artiodactyls. However, an IgG1 knockout cow or ewe has not been produced to experimentally test this assumption. The same is true in horses in which antibodies of different IgG subclasses have long been described, seven IgG subclasses are known from gene sequences but evidence in support of a unique role for each of these subclass antibodies in the horse immune response is lacking. The study of the many IgG subclasses in swine is only in its infancy because until recently, the subclasses had not been defined (Butler et al., 2009c; Kloep et al., 2012). This lack of progress has been largely due to the lack of IgG subclass reagents for use in immunoassays and, of course, the lack of subclass knockout animals. A solution to the first problem is now underway (Butler et al., 2012a,b)

The point to be made is that with the possible exception of ruminant IgG1, evidence is not strong that subclass diversification of IgG is really necessary for survival of the host especially since rabbits accomplish this with one gene for IgG. Perhaps the IgA polygeny of rabbits compensates for this “deficiency,” but there are no data that address this point and evidence for a unique role for each of the 13 different IgA subclasses in rabbits is missing.

THE ORIGIN OF VH AND Cy POLYGENY

As early as 1932, gene duplication was discussed as an essential feature of the evolutionary process (Haldane, 1932; Bridges, 1936;

Ohno, 1970). Various mechanisms are known or proposed including RNA/DNA transposition, non-homologous crossing over, and even entire gene duplication (Woody and McConkey, 2011). The first two are often referred to as genomic gene conversion, although this term seems most appropriate to explain non-homologous crossing over (Meselson and Radding, 1975; Szostak et al., 1983). In humans, mouse and rat genomes, ~15% of all genes represent tandemly arrayed genes (Shoja and Zhang, 2006). Tandem duplication also creates redundancy (Li et al., 2005) and this is what is seen among Ig genes, especially those encoding the variable heavy and light chain loci and in some species the Cy genes. Tandem arrayed duplicons are quite often conserved among species (Zhang, 2003); a phenomenon that also appears true for the Ig variable region genes of swine (Eguchi-Ogawa et al., 2010). Since tandem duplicates are mainly attributed to non-homologous cross-overs, many or most emerge as functional genes (Woody and McConkey, 2011). During evolution these duplicons appear to be retained in the genome despite their apparent redundancy (Xue and Fu, 2009). Duplication reduced selective pressure on single genes since one or several of the duplicated “offspring” genes can more rapidly accumulate mutations and therefore assume new functions while not comprising the function of the parent gene (Ohno, 1970; Lynch and Conery, 2000; Kondrashov et al., 2002).

While the exact mechanism remains unknown, evidence points to non-homologous recombination (gene conversion), acting together with gene duplication, as the mechanism for VH and Cy polygeny in mammals. This is consistent with studies that some proportion of these duplicated genes will be non-functional, i.e., pseudogenes (Wolfe and Shields, 1997). This is specifically seen among the tandem duplicons in the VH, V κ , V λ loci, and among especially the Cy genes of the constant heavy chain sublocus in all higher vertebrates. Among the duplicated VH and Cy gene in swine, the 5' region that encodes FR1 or CH1 respectively, is most conserved among duplicons and among species. While this might suggest progressive 5' to 3' mutation of the duplicated genes during evolution, this does not seem to be the case for the VH and Cy genes of swine. For example, the hinge regions of Cy genes are most variable in swine and other mammals (Figure 9; Butler et al., 2009c). While the FR regions of porcine VH genes are nearly identical, the CDR1 and CDR2 segments differ and are shared. Thus the CDR regions of VH genes and the hinge segment of Cy genes are perhaps the best candidates for genomic gene conversion rather than point mutation during evolution.

SHM REDUCES THE VALUE OF VARIABLE REGION GENE POLYGENY

Most evolutionary genetic studies of protocaryotic and eucaryotic species focus on changes in the genome that are transmitted to the offspring in Mendelian fashion. Studies on adaptive immunity, a system most developed in mammals, introduced a new mechanism for variability, namely somatic generation of antibody specificity. This process involves somatic rearrangement of gene segments, additions/subtractions of nucleotide at the boundaries of rearranged segments plus SHM of these rearrangements. The rate of the latter is several logs greater than for the mutations that accumulate in the genome.

Well-developed adaptive immune systems that follow this pattern are primarily restricted to mammals and the chicken and are

associated with the expression of AID (see below). Such somatic processes are difficult to identify in cartilaginous and bony fishes and are present at reduced frequency in amphibians (Du Pasquier et al., 1998, 2000). At least SHM and CSR in mammals are associated with the formation of germinal centers that involves the expression of a member of the APOBEC family called antigen-activated cytidine deaminase (AID; Honjo et al., 2002). In lieu of this feature, elasmobranches have ~200 cassettes of fused V-D-J-C genes in their genome although there are some variations on this theme (Dooley and Flajnik, 2006). The genome of *Xenopus tropicalis* contains 11 VH gene families and 37 V λ genes encoded at three loci (Qin et al., 2008). Among eutherian mammals, some bats have >250 VH genes, lab rodents and human ~100, and rabbits ~200 (Bratsch et al., 2011; **Table 1**). Many of these VH genes are known to be pseudogenes, consistent with data on tandem gene duplication (Wolfe and Shields, 1997). In general, bats (Chiroptera) and rodents which are considered primitive eutherian mammals, have more VH genes whereas Ungulates, that emerged later, have few VH genes (**Table 1**). This is superficial support for the hypothesis that the generation of antibody diversity by SHM or somatic gene conversion has minimized the need for large numbers of variable region gene segments and the importance of combinatorial diversity in more recently evolved mammals.

We believe that vertebrates initially evolved the need for multiple VH genes to create a repertoire of specific antibodies but with the subsequent development of somatic rearrangement and especially SHM, the needs for such polygeny became increasingly redundant. Thus, tandemly duplicated VH genes remain in the genome as evolutionary relics.

LESSONS FROM STUDIES ON ANTIBODY REPERTOIRE DEVELOPMENT IN PIGLETS

The use of the piglet model to study antibody repertoire development has provided useful information on the B cell system of this

REFERENCES

- Acha-Orbea, H., Finke, D., Attinger, A., Schmid, S., Wehrli, N., Vacheron, S., Xenarios, I., Scarpellino, L., Toellner, K. M., MacLennan, I. C., and Lutter, S. A. (1999). Interplay between mouse mammary tumor virus and the cellular and humoral immune response. *Immunol. Rev.* 168, 287–303.
- Allan, G. M., and Ellis, J. A. (2000). Porcine circoviruses: a review. *J. Vet. Diagn. Invest.* 12, 3–14.
- Becker, R. S., and Knight, K. L. (1990). Somatic diversification of immunoglobulin heavy chain VDJ genes: evidence for somatic gene conversion in rabbit. *Cell* 63, 987–997.
- Berek, C., and Milstein, C. (1987). Mutation drift and repertoire shift in the maturation of the immune response. *Immunol. Rev.* 96, 23–41.
- Brambell, F. W. R. (1970). *The Transmission of Passive immunity from Mother to Young*. Amsterdam: North-Holland Publishing Company.
- Bratsch, S., Wertz, N., Chaloner, K., Kunz, T. H., and Butler, J. E. (2011). The little brown bat displays a highly diverse VH, DH and JH repertoire but little evidence of somatic hypermutation. *Dev. Comp. Immunol.* 35, 421–430.
- Bridges, C. B. (1936). The bar “gene” a duplication. *Science* 83, 210–211.
- Bridges, S. L. Jr., Lee, S. K., Johnson, M. L., Lavelle, J. C., Fowler, P. G., Koopman, W. J., and Schroeder, H. W. Jr. (1995). Somatic mutation and CDR3 lengths of immunoglobulin kappa light chains expressed in patients with rheumatoid arthritis and in normal individuals. *J. Clin. Invest.* 96, 831–841.
- Burnett, R. C., Hanly, W. C., Zhai, S. K., and Knight, K. L. (1989). The IgA heavy chain gene family in rabbits: cloning and sequence analysis of 13 Ca genes. *EMBO J.* 8, 40410–40407.
- Butler, J. E. (1974). “Immunoglobulins of the mammary secretions” in *Lactation, a Comprehensive Treatise*, Vol. III, Chap. V, eds B. L. Larson and V. Smith (New York: Academic Press), 217–255.
- Butler, J. E. (1983). Bovine immunoglobulins: an augmented review. *Vet. Immunol. Immunopathol.* 4, 43–152.
- Butler, J. E. (1997). Immunoglobulin gene organization and the mechanism of repertoire development. *Scand. J. Immunol.* 45, 455–462.
- Butler, J. E., and Kehrl, M. E. Jr. (2005). “Immunocytes and immunoglobulins in milk,” in *Mucosal Immunology*, 3rd Edn, eds J. Mestecky, M. E. Lamm, W. Strober, J. R. McGhee, L. Mayer, and J. Bienenstock (New York: Academic Press), 1763–1793.
- Butler, J. E., Lager, K. M., Splichal, I., Francis, D., Kacskovics, I., Sinkora, M., Wertz, N., Sun, J., Zhao, Y., Brown, W. R., DeWald, R., Dierks, S., Muyldermans, S., Lunney, J. K., McCray, P. B., Rogers, C. S., Welsh, M. J., Navarro, P., Klobasa, F., Habe, F., and Ramsoondar, J. (2009a). The piglet as a model for B cell and immune system development. *Vet. Immunol. Immunopathol.* 128, 147–170.
- Butler, J. E., Sinkora, M., Wertz, N., and Kacskovics, I. (2009b). Immunoglobulins, B cells and repertoire development. *Dev. Comp. Immunol.* 33, 321–333.
- Butler, J. E., Wertz, N., Deschacht, N., and Kacskovics, I. (2009c). Porcine IgG: structure, genetics and evolution. *Immunogenetics* 61, 209–230.
- Butler, J. E., Lemke, C. D., Weber, P., Sinkora, M., and Lager, K. D. (2007). Antibody repertoire development in fetal and neonatal piglets. XIX. Undiversified B cells with hydrophobic HCDR3s preferentially proliferate in PRRS. *J. Immunol.* 178, 6320–6331.

species that often differs from the textbook models describing the process. Thus, the basis of the title of the review. It is also inconsistent with proposals that place hoofed mammals in the GALT category in which development of the B cell repertoire depends on hindgut lymphoid tissue (Lanning et al., 2004). Collectively considered, our studies indicate that:

- (1) Diversity of the pre-immune repertoire in swine is almost exclusively dependent on junctional diversity in CDR3 of the heavy chain followed by SHM, since only seven VH genes with shared CDRs, two DH genes and one functional JH comprise the functional repertoire and such diversity is greatly restricted in light chain rearrangements.
- (2) Repertoire diversification after antigen encounter is by SHM primarily in the CDR regions of the rearranged heavy chain variable regions. Evidence for somatic gene conversion, junctional diversity or SHM in light chains that might contribute to extensive antibody repertoire diversification is lacking.
- (3) Light chain rearrangement occurs first in the lambda locus and there is no evidence of a special V λ gene like λ 5 that is used in the earliest phases of B cell lymphogenesis. This raises the question of whether the conventional pre-BCR is present in swine or any artiodactyls.
- (4) The prominent hindgut lymphoid tissue of swine, the IPP, is not required for B cell lymphogenesis, maintenance of B cell levels or repertoire development. This opens the question regarding the true function of this lymphoid organ. Here we propose it represents “first responder” mucosal lymphoid tissue. A key factor in this scenario is that IgG3, which is encoded by the most 5' Cy gene, shares this same genomic feature with mice, cattle, and humans. Perhaps this primordial IgG provides the “natural antibodies” that target the polysaccharide antigens of intestinal bacteria.

- Butler, J. E., Sun, J., and Navarro, P. (1996). The swine immunoglobulin heavy chain locus has a single JH and no identifiable IgD. *Int. Immunol.* 8, 1897–1904.
- Butler, J. E., Sun, J., Weber, P., Ford, S. P., Rehakova, Z., Sinkora, J., and Lager, K. (2001). Antibody repertoire development in fetal and neonatal piglets. IV. Switch recombination, primarily in fetal thymus occurs independent of environmental antigen and is only weakly associated with repertoire diversification. *J. Immunol.* 167, 3239–3249.
- Butler, J. E., Sun, X.-Z., Wertz, N., Lager, K. M., Urban, J. Jr., Nara, P., and Tobin, G. (2011a). Antibody repertoire development in fetal and neonatal piglets. XXI. VH usage remains constant in fetal piglets and postnatally development. *Mol. Immunol.* 49, 483–494.
- Butler, J. E., Mateo, K., Sun, X.-Z., Wertz, N., Sinkora, M., Harvey, R., and Francis, D. L. (2011b). Antibody repertoire development in fetal and neonatal piglets XX: the ileal Peyer's patches are not a site of B cell lymphogenesis and are not required for systemic B cell proliferation and Ig synthesis. *J. Immunol.* 187, 5141–5149.
- Butler, J. E., Sun, X.-Z., and Wertz, N. (2011c). “Immunoglobulin polygeny: an evolutionary perspective” in *Gene Duplication*, ed. F. Friedberg (Rijeka: InTech), 113–140.
- Butler, J. E., Wertz, N., Zhao, Y., Kunz, T. H., Bratsch, S., Whitaker, J., and Schountz, T. (2011d). Two suborders of bats have the canonical isotypes repertoire of other eutherian mammals. *Dev. Commun. Immunol.* 35, 272–284.
- Butler, J. E., Sun, X.-Z., Wertz, N., Vincent, A. L., Zanella, E. L., and Lager, K. M. (2012a). Antibody repertoire development in fetal and neonatal piglets. XVI. Influenza stimulates adaptive immunity, class switch and diversification of the IgG repertoire encoded by downstream Cγ genes. *Immunology* (in press).
- Butler, J. E., Wertz, N., Sun, X.-Z., Lunney, J. K., and Muyldermanns, J. (2012b). Resolution of an immunodiagnostic dilemma: heavy chain chimeric antibodies for species in which plasmocytomas are unknown. *Mol. Immunol.* (pending).
- Butler, J. E., Weber, P., Sinkora, M., Baker, D., Schoenherr, A., Mayer, B., and Francis, D. (2002). Antibody repertoire development in fetal and neonatal piglets. VIII. Colonization is required for newborn piglets to make serum antibodies to T-dependent and type 2 T-independent antigens. *J. Immunol.* 169, 6822–6830.
- Butler, J. E., Weber, P., Sinkora, M., Sun, J., Ford, S. J., and Christenson, R. (2000a). Antibody repertoire development in fetal and neonatal piglets. II. Characterization of heavy chain CDR3 diversity in the developing fetus. *J. Immunol.* 165, 6999–7011.
- Butler, J. E., Sun, J., Weber, P., and Francis, D. (2000b). Antibody repertoire development in fetal and neonatal pigs. III. Colonization of the gastrointestinal tracts results in preferential diversification of the pre-immune mucosal B-cell repertoire. *Immunology* 100, 119–130.
- Butler, J. E., Weber, P., and Wertz, N. (2006a). Antibody repertoire development in fetal and neonatal pigs. XIII. “Hybrid VH genes” and the pre-immune repertoire revisited. *J. Immunol.* 177, 5459–5470.
- Butler, J. E., Sinkora, M., Wertz, N., Holtmeier, W., and Lemke, C. D. (2006b). Development of the neonatal B- and T-cell repertoire in swine: implications for comparative and veterinary immunology. *Vet. Res.* 37, 417–441.
- Butler, J. E., Weber, P., Wertz, N., and Lager, K. M. (2008). Porcine reproductive and respiratory syndrome virus (PRRSV) subverts development of adaptive immunity by proliferation of germline-encoded B cells with hydrophobic HCDR3s. *J. Immunol.* 180, 2347–2356.
- Butler, J. E., and Wertz, N. (2006). Antibody repertoire development in fetal and neonatal piglets. XVII. IgG subclass transcription revisited with emphasis on new IgG3. *J. Immunol.* 177, 5480–5489.
- Butler, J. E., Wertz, N., Sun, J., Wang, H., Lemke, C., Chardon, P., Puimi, F., and Wells, K. (2005a). The pre-immune variable kappa repertoire of swine is selectively generated from certain subfamilies of $\text{V}\kappa 2$ and one $\text{J}\kappa$ gene. *Vet. Immunopathol.* 108, 127–137.
- Butler, J. E., Francis, D., Freeling, J., Weber, P., Sun, J., and Krieg, A. M. (2005b). Antibody repertoire development in fetal and neonatal piglets. IX. Three PAMPs act synergistically to allow germfree piglets to respond to TI-2 and TD antigens. *J. Immunol.* 175, 6772–6785.
- Butler, J. E., Wertz, N., Wang, H., Sun, J., Chardon, P., Piumi, F., and Wells, K. (2004). Antibody repertoire in fetal and neonatal pigs. VII. Characterization of the pre-immune kappa light chain repertoire. *J. Immunol.* 173, 6794–6805.
- Cohn, M., and Langman, R. E. (1990). The protector: the unit of humoral immunity selected by evolution. *Immunol. Rev.* 115, 11–147.
- Coutelier, J.-P., Coulie, G., Wauters, P., Heremans, H., and der Logt, J. T. (1990). In vivo polyclonal B-lymphocyte activation elicited by murine viruses. *J. Virol.* 64, 5383–5388.
- Cunnington, P. G., and Naysmith, J. D. (1975). Naturally occurring double-stranded RNA and immune responses. III. Immunogenicity and antigenicity in animals. *Immunology* 29, 1001–1017.
- Dooley, H., and Flajnik, M. F. (2006). Antibody repertoire development in cartilaginous fish. *Dev. Comp. Immunol.* 30, 43–56.
- Du Pasquier, L., Robert, J., Courtelet, M., and Mussmann, R. (2000). B cell development in the amphibian *Xenopus*. *Immunol. Rev.* 175, 201–213.
- Du Pasquier, L., Wilson, M., Greenberg, A. S., and Flajnik, M. F. (1998). Somatic mutation in ectothermic vertebrates: musings on selection and origins. *Curr. Top. Microbiol. Immunol.* 229, 199–216.
- Eguchi-Ogawa, T., Sun, X.-Z., Wertz, N., Uenishi, H., Puimi, F., Chardon, P., Wells, K., Tobin, G. J., and Butler, J. E. (2010). Antibody repertoire development in fetal and neonatal piglets. XI. The relationship of VDJ usage and the genomic organization of the variable heavy chain locus. *J. Immunol.* 184, 3734–3742.
- Eguchi-Ogawa, T., Toki, D., Wertz, N., Butler, J. E., and Uenishchi, H. (2012). Complete structure of the genomic sequence encoding the constant region of the porcine immunoglobulin heavy chain. *Mol. Immunol.* (in press).
- Ehrlich, R. (1995). Selective mechanisms utilized by persistent and oncogenic viruses to interfere with antigen producing and presentation. *Immunol. Res.* 14, 77–97.
- Girschick, H. J., and Lipsky, P. E. (2001). The kappa repertoire of human neonatal B cells. *Mol. Immunol.* 38, 1113–1127.
- Glas, A. M., van Monfort, E. H. N., and Milner, E. C. B. (2000). “The human antibody repertoire: old notions, current realities and VH gene-dependent biases,” in *The Antibodies*, Vol. 6, eds M. Zanetti and J. D. Capra (Amsterdam: Harwood Academic Publishers), 63–79.
- Hahn, G., Jores, R., and Mocarski, E. S. (1998). Cytomegalovirus remains latent in a common precursor of dendritic and myeloid cells. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3937–3942.
- Haldane, J. B. S. (1932). *The Causes of Evolution*. London: Longmans and Green.
- Hamers-Casterman, C., Atarhouch, T., Muyldermanns, S., Robinson, G., Hamers, C., Bajana Songa, E., Bendahman, N., and Hamers, R. (1993). Naturally occurring antibodies devoid of light chains. *Nature* 363, 446–448.
- Herzenberg, L. A., Stall, A. M., Lalor, P. A., Sidman, C., Moore, W. A., Parks, D., and Herzenberg, L. A. (1986). The Ly-1 B cell lineage. *Immunol. Rev.* 93, 81–102.
- Honjo, T., Kinoshita, K., and Muramatsu, M. (2002). Molecular mechanisms of lass switch recombination: linkage with somatic hypermutation. *Annu. Rev. Immunol.* 20, 165–196.
- Hood, L., Gray, W. R., Saunders, B. G., and Dreyer, W. J. (1967). Light chain evolution. *Cold Spring Harb. Symp. Quant. Biol.* 32, 133–146.
- Hunziker, L., Recher, M., Macpherson, A. J., Ciurea, A., Freigang, S., Hengartner, H., and Zinkernagel, R. M. (2003). Hypergammaglobulinemia and autoantibody induction mechanisms in viral infection. *Nat. Immunol.* 4, 343–349.
- Kehoe, J. M., and Capra, J. D. (1974). Nature and significance of immunoglobulin subclasses. *NY State J. Med.* 74, 489–491.
- Kloep, A., Wertz, N., Mendicino, M., and Butler, J. E. (2012). Linkage haplotype for IgG and IgA subclass genes. *Immunogenetics* 64, 469–473.
- Knight, K. L. (1992). Restrictive VH gene usage and generation of antibody diversity in rabbit. *Annu. Rev. Immunol.* 10, 593–616.
- Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biol.* 3, RESEARCH0008.
- Lanning, D., Osbourne, B. A., and Knight, K. L. (2004). “Immunoglobulin genes and generation of antibody repertoires in higher vertebrates: a key role for GALT,” in *Molecular Biology of B Cells*, eds F. W. Alt, T. Honjo, and M. S. Neuberger (London: Elsevier Science Ltd.), 433–448.
- Lanning, D. K., Zhao, S. K., and Knight, K. L. (2003). Analysis of the 3' Cmu region of the rabbit Ig heavy chain locus. *Gene* 309, 135–144.
- LeFranc, G., Chaabani, H., Van Loghem, E., Lefranc, M. P., De Lange, G., and Helal, A. N. (1983). Simultaneous absence of the human IgG1, IgG2, IgG4 and IgA1 subclasses:

- immunological and immunogenetic considerations. *Eur. J. Immunol.* 13, 240–244.
- LeFranc, M. P., Hammarstrom, L., Smith, C. I., and Lefranc, G. (1991). Gene deletion in the human immunoglobulin heavy chain constant region locus: molecular and immunological analysis. *Immunol. Rev.* 2, 265–281.
- Lemke, C. D., Haynes, J. S., Spaete, R., Adolphson, D., Vorwald, A., Lager, K., and Butler, J. E. (2004). Lymphoid hyperplasia resulting in immune dysregulation is caused by PRRSV infection in pigs. *J. Immunol.* 172, 1916–1925.
- Li, W. H., Yang, J., and Guo, X. (2005). Expression divergence between duplicate genes. *Trends Genet.* 21, 602–607.
- Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155.
- Mage, R. G., Lanning, D., and Knight, K. L. (2006). B cell and antibody repertoire development in rabbits: the requirement of gut-associated lymphoid tissues. *Dev. Comp. Immunol.* 30, 137–153.
- Merial, Inc. (2004). PCV2 diseases: from research back to the field again, Vol. 5. Kansas City, MO: Corporate Publication by Merial Inc.
- Meselson, M. S., and Radding, C. M. (1975). A general model for genetic recombination. *Proc. Natl. Acad. Sci. U.S.A.* 72, 358–361.
- Mossman, T. R., and Coffman, R. L. (1989). TH1 and TH2 cells: different patterns of lymphokine secretion lead to different functional properties. *Annu. Rev. Immunol.* 7, 143–173.
- Mullens, M. A., Register, K. B., Bayles, D. O., and Butler, J. E. (2011). *Haemophilus parasuis* exhibits IgA protease activity but lacks homologs of the IgA protease genes of *Haemophilus influenzae*. *Vet. Microbiol.* 153, 407–412.
- Navarro, P., Christenson, R., Weber, P., Rothschild, M., Ekhard, G., Lemky, J., and Butler, J. E. (2000). Porcine IgA allotypes are not equally transcribed or expressed in heterozygous swine. *Mol. Immunol.* 37, 653–664.
- Nguyen, V. K. (2001). *Generation of Heavy Chain Antibodies in Camelids*. Ph.D. thesis, Free University of Brussels, Brussels, 109–111.
- Nguyen, V. K., Su, C., Muylldermans, S., and van der Loo, W. (2002). Heavy chain antibodies in Camelids: a case of evolutionary innovation. *Immunogenetics* 54, 39–47.
- Nitschke, L., Kosco, M. L., Kohler, G., and Lamers, M. C. (1993). Immunoglobulin D deficient mice can mount normal immune responses to thymus-independent and -dependent antigens. *Proc. Natl. Acad. Sci. U.S.A.* 90, 1887–1891.
- Ochsenebein, A. F., and Zinkernagel, R. (2000). Natural antibodies and complement link innate and acquired immunity. *Immunol. Today* 1, 624–630.
- Ohno, S. (1970). *Evolution by Gene Duplication*. New York: Springer-Verlag.
- Olsson, P. G., Rabbani, H., Hammarstrom, L., and Smith, C. I. (1993). Novel human immunoglobulin heavy chain constant region gene deletion haplotypes characterized by pulsed-field electrophoresis. *Clin. Exp. Immunol.* 94, 84–90.
- Padlan, E. A. (1994). Anatomy of the antibody molecule. *Mol. Immunol.* 31, 169–217.
- Plaut, A. G., Wustar, R. Jr., and Capra, J. D. (1974). Differential susceptibility of human IgA immunoglobulins to streptococcal IgA proteases. *J. Clin. Invest.* 54, 1295–1300.
- Qin, T., Ren, L., Hu, X., Guo, Y., Fei, J., Pan-Hammarstrom, Q., Butler, J. E., Wu, C., Li, L., Hammarstrom, L., and Zhao, Y. (2008). Genomic organization of the immunoglobulin gene loci in *Xenopus tropicalis*: evolutionary implications. *Dev. Comp. Immunol.* 32, 156–165.
- Rabbani, H., Kondo, N., Smith, C. I., and Hammarstrom, L. (1995). The influence of gene deletion and duplication within the IgHC locus on serum immunoglobulin subclass levels. *Clin. Immunol. Immunopathol.* 76, 214–218.
- Rajewsky, K., Forster, I., and Cumano, A. (1987). Evolutionary and somatic selection of the antibody repertoire in the mouse. *Science* 238, 1088–1094.
- Ratcliffe, M. J. H. (2006). Antibodies, immunoglobulin genes and the bursa of Fabricius in chicken. B cell development. *Dev. Comp. Immunol.* 30, 101–118.
- Reynaud, C. A., Anquez, V., Daher, A., and Weill, J.-C. (1987). A hyperconversion mechanism generates the chicken pre-immune light chain repertoire. *Cell* 48, 379–388.
- Richl, P., Stern, U., Lipsky, P. E., and Girschick, H. J. (2008). The lambda gene immunoglobulin repertoire of human neonatal B cells. *Mol. Immunol.* 45, 320–327.
- Rodkey, L. S., and Adler, F. L. (1983). Regulation of natural anti-allotypic antibody responses by network induced auto-anti-idiotypic responsiveness of their offspring. *J. Exp. Med.* 152, 1024–1035.
- Schelonka, R. L., Tanner, J., Zhang, Y., Gartland, G. L., Zemlin, M., and Schroeder, H. W. Jr. (2007). Categorized selection of the antibody repertoire in splenic B cells. *Eur. J. Immunol.* 4, 1010–1021.
- Schiaffella, E., D. Sehgal, A.O., Anderson, and Mage, R. G. (1999). Gene conversion and hypermutation during diversification of VH sequences in developing germinal centers of immunized rabbits. *J. Immunol.* 162, 3984–3995.
- Schwartz, J. C., Lefranc, M., and Murtaugh, M. P. (2012a). Evolution of the porcine kappa locus through germline gene conversion. *Immunogenetics* 64, 303–311.
- Schwartz, J. C., Lefranc, M., and Murtaugh, M. P. (2012b). Organization, complexity and allelic diversity of the porcine immunoglobulin lambda locus. *Immunogenetics* 64, 399–407.
- Sheehan, K. M., Mainville, C. V. A., Willert, S., and Brodeur, P. H. (1993). The utilization of individual VH exons in the primary repertoire of adult BALB/c mice. *J. Immunol.* 151, 5363–5375.
- Shoja, V., and Zhang, L. (2006). A roadmap of tandemly arrayed genes in the genomes of human, mouse and rat. *Mol. Biol. Evol.* 23, 2134–2141.
- Sinkora, M., Sun, J., Sinkorova, J., Christenson, R. K., Ford, S. P., and Butler, J. E. (2003). Antibody repertoire development in fetal and neonatal piglets. VI. B cell lymphogenesis occurs in multiple sites with differences in the frequency of in-frame rearrangements. *J. Immunol.* 170, 1781–1788.
- Sinkora, M., Zelena, K., Butler, J. E., Francis, D., Santiago-Mateo, K., Potockova, H., and Sinkorova, J. (2011). Ileal Peyer's patches (IPP) are not necessary for B cell development and maintenance and do not contribute significantly to the overall B cell pool in swine. *J. Immunol.* 187, 5150–5161.
- Skvaril, F., Barandén, S., Kuffer, F., and Probst, M. (1976). Changes of kappa/lambda ratio of human serum immunoglobulins in the course of development. *Blut* 33, 281–284.
- Sun, J., and Butler, J. E. (1997). Sequence analysis of swine switch μ , $C\mu$ and $C\mu m$. *Immunogenetics* 46, 452–460.
- Sun, J., Hayward, C., Shinde, R., Christenson, R., Ford, S. P., and Butler, J. E. (1998). Antibody repertoire development in fetal and neonatal piglets. I. Four VH genes account for 80% of VH usage during 84 days of fetal life. *J. Immunol.* 161, 5070–5078.
- Sun, J., Kacskovics, I., Brown, W. R., and Butler, J. E. (1994). Expressed swine VH genes belong to a small VH gene family homologous to human VH III. *J. Immunol.* 153, 5618–5627.
- Sun, X.-Z., Wertz, N., Lager, K., Sinkora, M., Stepanova, K., Tobin, G., and Butler, J. E. (2012a). Antibody repertoire development in fetal and neonatal piglets. XXII. Lambda rearrangement precedes kappa rearrangement during B cell lymphogenesis in swine. *Immunol. (British)* (in press).
- Sun, X.-Z., Wertz, N., Lager, K. L., Tobin, G., and Butler, J. E. (2012b). Antibody repertoire development in fetal and neonatal piglets. XXIII: fetal piglets infected with a vaccine strain of PRRS Virus display the same immune dysregulation seen in isolator piglets. *Vaccine* 30, 3646–3652.
- Sun, X.-Z., Wertz, N., Lager, K. L., and Butler, J. E. (2012c). Antibody repertoire development in fetal and neonatal piglets. XV. Porcine circovirus type 2 infection results in serum IgG antibodies to ORF 2, elevated IgA levels but little evidence for immune suppression. *Vaccine* (pending).
- Szostak, J. W., Orr-Weaver, T. L., and Rothstein, R. J. (1983). The double-strand break repair model for recombination. *Cell* 33, 25–35.
- Tiegs, S. L., Russell, D. M., and Nemazee, D. (1993). Receptor editing in self-reactive bone marrow B cells. *J. Exp. Med.* 177, 1009–1020.
- Tristram, D. A. (2005). "Maternal genital tract infection and the neonates," in *Mucosal Immunology*, 3rd Edn, Vol II, eds J. Mestecky, M. E. Lamm, W. Strober, J. Bienestock, J. R. McGhee, and L. Mayer (Elsevier/Academic Press), 1721–1731.
- Vazquez, J., Wertz, N., Sun, J., Wells, K., Sun, X.-Z., and Butler, J. E. (2012). Antibody repertoire development in fetal and neonatal piglets. XI. Characterization of the expressed lambda repertoire. *Mol. Immunol.* (pending).
- Victor, K. D., Vu, K., and Feeney, A. J. (1994). Limited junctional diversity in kappa light chains. Junctional sequences from CD431B2201 early B

- cell progenitors resemble those from peripheral B cells. *J. Immunol.* 152, 3467–3475.
- Wang, H., and Shlomchik, M. J. (1998). Maternal Ig mediates neonatal tolerance in rheumatoid factor transgenic mice but tolerance breaks down in adult mice. *J. Immunol.* 160, 2263–2271.
- Weinstein, J. A., Jiang, N., White R. A. III, Fisher, D. S., and Quake, S. R. (2009). High throughput sequencing of the Zebrafish antibody repertoire. *Science* 324, 807–810.
- Wikler, M., Demeur, C., Dewasne, G. and Urbain, J. (1980). Immunoregulatory role of maternal idiotypes. Ontogeny of immune networks. *J. Exp. Med.* 152, 1024–1035.
- Winstead, C. R., Zhai, S. K., Sethupathi, P., and Knight, K. L. (1999). Antigen-induced somatic diversification of rabbit IgA genes: gene conversion and point mutation. *J. Immunol.* 162, 6602–6612.
- Wolfe, K. H., and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713.
- Woody, O. Z., and McConkey, B. J. (2011). “Detection and analysis of functional specialization in duplicated genes,” in *Gene Duplication*, ed. F. Friedberg (Croatia: InTeck Rijeka), 37–58.
- Xu, J. L., and Davis, M. M. (2000). Diversity in the CDR3 region of VH is sufficient for most antibody specificities. *Immunity* 13, 37–45.
- Xue, C., and Fu, Y. (2009). Preservation of duplicate genes by originalization. *Genetica* 136, 69–78.
- Yamaguchi, N., Shimizu, S., Hara, A., and Saito, T. (1983). The effector maternal antigenic stimulation upon the active immune responsiveness of their offspring. *Immunology* 50, 229–238.
- Zhang, J. (2003). Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18, 292–298.
- Zhao, Y., Pan-Hammarstrom, Q., Kacsikovics, I., and Hammarstrom, L. (2003). The porcine Ig delta gene: unique chimeric splicing of the first constant region domain in its heavy chain transcripts. *J. Immunol.* 171, 1312–1318.
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Received:** 13 February 2012; **paper pending published:** 21 March 2012; **accepted:** 24 May 2012; **published online:** 27 June 2012.
- Citation:** Butler JE and Wertz N (2012) The porcine antibody repertoire: variations on the textbook theme. *Front. Immunol.* 3:153. doi: 10.3389/fimmu.2012.00153
- This article was submitted to Frontiers in B Cell Biology, a specialty of Frontiers in Immunology.
- Copyright © 2012 Butler and Wertz. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



Fundamental roles of the innate-like repertoire of natural antibodies in immune homeostasis

Jaya Vas¹, Caroline Grönwall¹ and Gregg J. Silverman^{1,2*}

¹ Laboratory of B Cell Immunobiology, Department of Medicine, New York University School of Medicine, New York, NY, USA

² Laboratory of B Cell Immunobiology, Department of Pathology, New York University School of Medicine, New York, NY, USA

Edited by:

Harry W. Schroeder, University of Alabama at Birmingham, USA

Reviewed by:

Harry W. Schroeder, University of Alabama at Birmingham, USA

David Nemazee, The Scripps Research Institute, USA

***Correspondence:**

Gregg J. Silverman, Laboratory of B Cell Immunobiology, Department of Medicine, New York University School of Medicine, Alexandria Center for Life Science, 450 East 29th Street, Room 804, New York, NY 10016, USA.
e-mail: gregg.silverman@nyumc.org

The composition of the early immune repertoire is biased with prominent expression of spontaneously arising B cell clones that produce IgM with recurrent and often autoreactive binding specificities. Amongst these naturally arising antibodies (NAbs) are IgM antibodies that specifically recognized damaged and senescent cells, often via oxidation-associated neo-determinants. These NAbs are present from birth and can be further boosted by apoptotic cell challenge. Recent studies have shown that IgM NAb to apoptotic cells can enhance phagocytic clearance, as well as suppress proinflammatory responses induced via Toll-like receptors, and block pathogenic IgG-immune complex (IC)-mediated inflammatory responses. Specific antibody effector functions appear to be involved, as these anti-inflammatory properties are dependent on IgM-mediated recruitment of the early recognition factors of complement. Clinical surveys have suggested that anti-apoptotic cell (AC) IgM NAbs may modulate disease activity in some patients with autoimmune disease. In mechanistic studies, anti-AC NAbs were shown to act in dendritic cells by inhibition of the mitogen-activated protein kinase (MAPK) pathway, a primary signal transduction pathway that controls inflammatory responses. This immunomodulatory pathway has an absolute requirement for the induction of MAPK phosphatase-1. Taken together, recent studies have elucidated the novel properties of a class of protective NAbs, which may directly blunt inflammatory responses through a primitive pathway for regulation of the innate immune system.

Keywords: immunoregulation, innate-like, natural antibody

INTRODUCTION

The evolutionary emergence of the combinatorial antigen receptor system of variable region (V) gene rearrangements in lymphocytes has provided a greatly enhanced capacity for specific recognition of an immense range of ligands. In humans, the antibody system can generate B cell antigen receptors (BCR) encoded by more than 10^{20} genetically distinct variable region rearrangements. Considering that each individual has only about 10^{11} lymphocytes, if the primary B cell repertoire was indeed generated randomly, there would be little or no recurrence in the antibody gene sequences in the somatically generated repertoires of the billions of humans on the planet. In the following sections, we review evidence the B cell compartment arises during development with a restricted and biased repertoire, and that the antibody products of these B cell clones may serve to protect the host from both external threats and for the maintenance of internal homeostasis.

RESTRICTION IN THE EARLY REPERTOIRE

In the mouse, there is a remarkable restriction in the usage patterns of the heavy chain V region (V_H) genes during early repertoire development (Perlmutter et al., 1985). Surveys of murine antibody sequences have provided extensive evidence of recurrent lymphocyte clones with the same V gene rearrangements in different individuals (Seidl et al., 1997), and emerging data suggests there

may be similar patterns in the human B cell repertoire (Jackson et al., 2012). In fact, these biases in the expression of V_H rearrangements are first detectable at a time point at which representation cannot be affected by antigenic selection of these IgM-associated clones (Schroeder et al., 1987; Schroeder and Wang, 1990). Moreover, a recent report suggested that the perinatal V_H repertoire expressed in human IgA may be even more restricted than the IgM pool (Rogosch et al., 2012). Recurrent biases in the V_H expression in the early B cell repertoire have also been reported in other species, such as swine (Sun et al., 1998) and sheep (Jenne et al., 2006), as well as more primitive species, such as the amphibian *Xenopus laevis* (Flajnik and Rumfelt, 2000), and zebrafish (Du Pasquier et al., 2000).

Both humans and mice have circulating IgM antibodies that arise early life without immunogenic challenge and have therefore been termed natural antibodies (NAbs). In fact, neonatal B cells produce IgM antibodies that are readily detectable in the bloodstream at birth, and studies in mice have shown that more than 80% of circulating IgM are produced by a phenotypically distinct mature B cell subset, termed the B-1a cell subset, and characterized by membrane-associated CD5. In general, while some B-1 cells express antigen-receptors for recognition of common bacterial Ags, some B-1 cell clones can also recognize self-antigens, including the phospholipidphosphatidylcholine (PtC), the phospholipid-associated phosphorylcholine (PC) head

group, as well as DNA and certain cell membrane proteins (Kantor and Herzenberg, 1993).

B-1 cells are believed to represent a developmentally distinct lineage from their adult counterpart, the bone marrow-derived B-2 subset (reviewed in Hardy, 2006; Baumgarth, 2011). Murine B-1 clones are self-replenishing, which ensures the maintenance of this repertoire, as later in life the capacity for *de novo* generation of mature lymphocytes with the B-1 cell phenotype is limited. Studies by Notkins and colleagues have shown that CD5-bearing human B cells also have a bias toward the production of certain types of autoantibodies (Casali and Notkins, 1989). However, CD5 molecules can represent an activation marker on human B cells, and hence by itself CD5 may not be a rigorous phenotypic marker for this B cell subset in humans (Cong et al., 1991). To address this long standing issue, Rothstein and coworkers have reported a detailed phenotyping scheme, in addition to CD5, for identifying human B cells with the diagnostic features of B-1 cells. The repertoire of these human B-1 cells also appeared to include prominent expression of self-specificities for native DNA and PC-containing antigens (Griffin et al., 2011).

AUTOREACTIVITY OF B LYMPHOCYTE SUBSETS

In mice, mature B-1 and B-2 lymphocyte subsets can play discrete but complementary functional roles in host defenses (reviewed in Baumgarth, 2011). There are also subpopulations within B-1 cells in addition to CD5⁺ B-1a cells, as B-1b cells (that do not express CD5) make essential contributions to T cell-independent defenses for certain types of infections (Alugupalli and Abraham, 2009). The clonal selection of these distinct B cell subsets may in part reflect differences in their cellular thresholds for negative selection (i.e., BCR-induced cell death) and in their activation requirements for second signals after BCR stimulation. By one estimate, over 70% of BCR-expressing immature B cells in the bone marrow display some level of autoreactivity while the level is much less in recirculating naïve mature B-2 cells (Wardemann et al., 2003). Hence, the immune tolerance checkpoints for B-2 cells that arise from precursors in the bone marrow appear to be generally more stringent in the removal of self-reactivity (i.e., negative selection). In contrast, conserved B-1 cell clonotypes may be positively selected (i.e., enhanced survival and clonal proliferation) by certain types of non-protein self-antigens (Hayakawa et al., 1999), which may include specific types of intracellular antigens (Ferry et al., 2007). As B-1 cells are a major source of circulating IgM in neonates, this may explain why neonatal IgM–Nabs from umbilical cord commonly display features of self-reactivity (Chou et al., 2009).

In the human immune system there is a remarkably strong association between the immune recognition of cell surface N-acetyllactosamine/polylactosamine determinants in glycoconjugates and the usage of the V_H4-34 gene segment (originally termed V_H4-21; Silberstein et al., 1991). N-acetyllactosamine/polylactosamine moieties are common on cell surfaces throughout the body, as these are structural components of the I/i blood group antigens and also constitute the antigenic target of pathogenic autoantibodies in cold-agglutinin disease (Silverman et al., 1990; Silberstein et al., 1991; Grillot-Courvalin et al., 1992; Pascual and Capra, 1992).

The immune recognition of I/i related non-protein antigens may be involved in very different types of immune responses. Using a lectin microarray system, exosomes released by human tumor cell lines were shown to express a shared polylactosamine glycan signature (Batista et al., 2011). These findings extend earlier evidence that I/i related determinants can be preferentially expressed on cells during early development and on their malignant counterparts (i.e., onco-fetal antigens; Feizi, 1988). In addition, V_H4-34 encoded autoantibodies were found to commonly bind to HIV-1 envelope determinants (Kobie et al., 2012). Batista et al. (2011) have suggested that these glycans reflect a recurrent type of glycan epitope profile on stressed and apoptotic cells (ACs). Taken together, these findings highlight the intertwined nature of immunodominant determinants on microparticles, exosomes, and HIV-1 virions that arise by budding through the membranes of stressed host cells.

B cell receptor encoded by V_H4-34 rearrangements recognize I/i determinants via contact sites associated with a V_H germline framework subdomain sequences – there is little apparent contribution from the somatically generated heavy chain CDR3 or by the paired light chain (Pascual et al., 1991). Using the 9G4 anti-idiotypic antibody, B cells that bear non-mutated V_H4-34 products have been shown to be highly represented, and whereas in healthy individuals these autoreactive germline B cells were shown to be excluded from T cell-dependent germinal center reactions (Pugh-Bernard et al., 2001), these V_H defined clones can be actively recruited into the germinal center reactions in patients with systemic lupus erythematosus (SLE; Cappione et al., 2005). These findings have been interpreted as evidence of immune defects in SLE patients related to the regulation of autoreactive B cells, although this topic remains controversial.

NATURAL ANTIBODIES AND IMMUNE RECOGNITION OF DAMAGED AND APOPTOTIC CELLS

During the process of AC death, different cell membrane-associated phospholipids can undergo selective enzyme-mediated and oxidation associated modifications, and these cell membrane neo-determinants become available for recognition by the immune system. Among these, phosphatidylserine (PS) becomes oxidized and rapidly translocates from the inner to the outer leaflet of the cell membrane upon the initiation of apoptosis, where it can serve as a recognition signal for ingestion by professional phagocytes (i.e., “eat me” signal). Apoptosis can also be associated with other lipid neo-determinants, such as malondialdehyde (MDA), which is formed from interactions of unsaturated lipids with reactive oxidation species. Oxidative modifications of the abundantly distributed neutral phospholipid, PtC, also affect the distribution and/or conformation of the PC head group (Friedman et al., 2002), which renders it accessible for antibody recognition. These PC-antigens, as well as MDA-containing antigens, are immunodominant within murine B cell clonal responses that are boosted by intravenous infusions of ACs (Chen et al., 2009b).

The dominant natural antibody-producing anti-PC B cell clone, termed T15 (also TEPC15), has recurrently been isolated in anti-PC responses. The high representation of this clone in the early repertoire in part reflects a bias for increased representation of the specific V_HS107.1 gene rearrangements used by the

T15 clone (Feeney, 1991). In fact both of the V_H and V_L rearrangements in the T15 clone are formed by primary sequence direct rearrangements, and are without somatic mutations. The invariance of the T15 clonotypic NAb therefore is reminiscent of germline encoded receptors of the innate immune system (discussed in Shaw et al., 2000). In fact, T15 clonotypic antibodies are highly specific for PC determinants (Kearney et al., 1981) and in microarray analysis demonstrated little or no cross-reactivity to a large number of structurally distinct antigens (Chen et al., 2009b). Throughout life, T15-related B-1 cells are a major source of NAbs to a range of PC-containing antigens (Masmoudi et al., 1990), including those present on AC membranes, oxidized low-density lipoprotein (LDL), as well as in pneumococcal bacterial cell wall polysaccharide (Shaw et al., 2003; Chou et al., 2009). Many other B-1 cell clones have been demonstrated to be polyreactive and relatively low-affinity (Kantor and Herzenberg, 1993). However, crystallographic analysis of a V_H S107.1 encoded (i.e., T15-related) Fab revealed a deep antigen-binding cleft with substantial binding affinity for the small PC moiety (reviewed in Davies et al., 1975). As a consequence of the dependence of the *in vivo* anti-PC response on T15 clonotypic B cells, otherwise immunocompetent mice which were made deficient only for the S107.1 V_H gene segment, have highly impaired responses to immune challenge with PC determinants on either ACs or bacteria, and also display impaired immune defenses for *S. pneumoniae* infection (Mi et al., 2000; Chen et al., 2009b). Taken together, these studies suggest that the antigen binding sites of T15-related antibodies have innate-like properties for recognition of PC-containing antigens are highly represented in the pre-immune repertoire (Kearney, 2005), in part because of their preferentially formation by biases in the somatic diversification mechanisms (Feeney, 1992).

Within the NAb pool there are also other antibodies that recognize distinct sets of neo-determinants that arise following cellular injury. There are at least two self-antigen specificities that have been reported to be associated with post-ischemic injury of endothelial cells (Zhang et al., 2008; Kulik et al., 2009). In addition, there are IgM-NAbs that specifically recognize erythrocytes with cell membrane-associated changes due to senescence or from damage by experimental treatment with the protease, bromalein (Cox and Hardy, 1985; Mercolino et al., 1986; Hardy and Hayakawa, 2005). These antibodies are reported to recognize a determinant involving PtC, although it is unclear whether the accessibility of this PtC-associated red cell epitope results from oxidative modification, or due to loss of erythrocyte membrane proteins. Red cell membrane intrinsic proteins have also been implicated as antigenic targets for IgG-NAbs (Lutz, 2012). Notably, as red cells have neither mitochondria nor nuclei, these cells do not undergo the same apoptosis-associated metabolic changes seen in conventional mitochondria-containing cells. This may explain why PC-related antigens are not prominently displayed on erythrocytes as a consequence of aging. Taken together, the cumulative data suggest that the IgM repertoire may include a range of distinct subsets of autoreactive NAbs, which recognize different cell types affected by apoptosis, injury and senescence, and these NAbs may help to regulate the clearance of different cell types and tissue remodeling as well as modulate innate immune responses.

EFFECTS OF PC-SPECIFIC NATURAL IgM ON TOLL-LIKE RECEPTOR-INDUCED INFLAMMATION

Earlier studies have shown that C1q can directly bind to AC membranes and then serve as an “eat-me” signal for the phagocytic clearance of these dying cells (Korb and Ahearn, 1997; Navratil et al., 2001; Ogden et al., 2001). In explanation, C1q may directly interact with externalized PS on these damaged cells (Paidassi et al., 2008). In some settings, the deposition of C1q onto ACs can subsequently have an immunomodulatory effect and inhibit the secretion of proinflammatory cytokines, although by itself these effects are limited (Fraser et al., 2009). Similar properties have also been associated with the mannose-binding lectin (MBL), which triggers the lectin pathway of complement activation. MBL is structurally related to C1q and these two recognition molecules share a common ancestral genetic origin (Matsushita et al., 2004). Furthermore, MBL can also bind directly to ACs. This may suggest that initiation of apoptosis is associated with a change in the distribution of high-mannose glycoconjugates on the cell membrane (Stuart et al., 2005). These findings are consistent with reports that phagocytes of C1q-deficient mice, as well as MBL-deficient mice, display defects in AC clearance (Quartier et al., 2005; Stuart et al., 2005).

The potential roles of T15 IgM-NAb have been investigated in the innate immune responses of professional phagocytes. As mentioned above, while this IgM natural antibody does not bind healthy cells it can specifically recognize exposed PC determinants on ACs and form complexes (Chen et al., 2009a,b). In turn, these AC-IgM complexes have greatly enhanced capacity to recruit the early complement factors, C1q and the structurally related MBL, at levels several-fold higher than in the absence of bound IgM. Notably the recruitment of C1q or MBL by IgM-NAb complexes greatly amplifies the capacity for AC phagocytic clearance (Chen et al., 2009a,b; illustrated in Figure 1). These properties are explained by reports that some polymeric IgM, when bound to their cognate antigen, are highly efficient at recruitment of C1q, while other studies have shown that polymeric IgM themselves can contain high mannose glycoconjugates (Arnold et al., 2006). Hence, AC-reactive polymeric IgM may serve to integrate these complement associated innate immune functions (Quartier et al., 2005; Chen et al., 2009a,b).

The formation of IgM-NAb complexes with ACs can also result in strong suppression of *in vivo* and *in vitro* inflammatory responses, including those induced by ligands for both membrane-associated and endosomal Toll-like receptors (TLRs), which include TLR3, TLR4, TLR7, and TLR9 (Chen et al., 2009b). These activities are also dependent on the recruitment of C1q and MBL, which are postulated to serve as bridging molecules that trigger phagocyte functions in a way that does not require activation of the complement cascade (Chen et al., 2009a,b). Hence, both the enhancement of apoptotic clearance and the down-modulation of inflammatory responses are therefore pathways by which some NAb may augment and amplify housekeeping functions that serve to protect the host.

EFFECTS OF THE APOPTOTIC CELL-SPECIFIC NAb–IgM ON IMMUNE COMPLEX DRIVEN PATHOGENESIS

During autoimmune pathogenesis high-affinity IgG autoantibodies can make direct contributions by multiple mechanisms

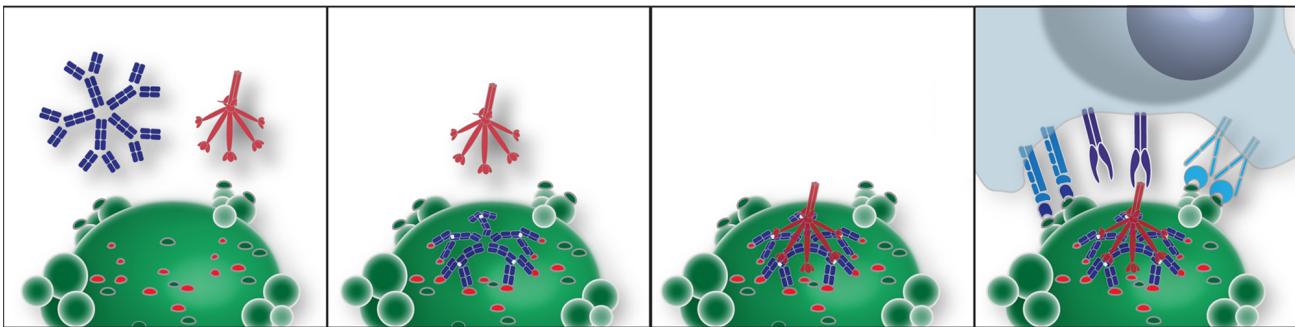


FIGURE 1 | Model of an IgM–NAB complex that enhances interactions between an apoptotic cell and professional phagocytes. In this idealized model, apoptotic death results in membrane alterations that expose a range of neo-determinants. PC-associated membrane determinants are recognized by antigen-binding sites of a pentameric IgM. Binding of PC determinants results in conformational changes in the mu constant regions, which expose a

conformational site responsible for recruitment of the globular heads of C1q (Czajkowsky and Shao, 2009). Alternatively, MBL binds to nearby high mannose N-linked glycoconjugates on mu-associated sites (not shown). This polymeric IgM–C1q complex is involved in generating or stabilizing interactions with receptors on professional phagocytes, which enhance apoptotic phagocytosis and blocks inflammatory responses.

(reviewed in Elkon and Casali, 2008; Lleo et al., 2010). Central to these pathways, many IgG autoantibodies implicated in systemic autoimmune diseases form immune complexes (ICs) with their target antigen, which alter their potential interactions with the host immune system. In sera of patients with conditions such as SLE and Sjögren's syndrome, nucleic acid-recognizing autoantibodies can form ICs with their antigens, ribonucleoproteins, or deoxyribonucleoproteins. These ICs can deposit in tissues such as the kidney, skin, and joints and drive local inflammation and tissue injury by triggering complement cascades (reviewed in Bagavant and Fu, 2009). IgG-ICs might additionally serve as a delivery system to transport self-antigens to endosomal pattern-recognition receptors (PRRs) via uptake by activating Fc receptors. Studies by Marshak-Rothstein, Rifkin, and Rönnblom have demonstrated that DNA or RNA-containing ICs consisting of IgG autoantibodies bound to nuclear debris from dying cells can activate mouse B cells (Leadbetter et al., 2002), conventional dendritic cells (DCs; Boulé et al., 2004), and plasmacytoid DCs (Yasuda et al., 2007) as well as human peripheral blood mononuclear cells (Lövgren et al., 2004). These IgG-nucleic acid ICs can interact with rheumatoid factor autoantibody-bearing B cells (Leadbetter et al., 2002) or with DCs bearing activating Fc γ R, which enables the delivery to otherwise inaccessible intracellular compartments. The outcome of this targeted DNA/RNA internalization is the activation of PRRs of the innate immune system, such as the nucleic acid-recognizing TLRs; TLR7 (activated by ssRNA) and TLR8 and TLR9 (activated by DNA). IgG antibodies to the citrullinated self-protein, fibrinogen, which are prevalent in rheumatoid arthritis patients, have also been shown *in vitro* to activate macrophages through a co-stimulatory pathway involving both Fc γ R and TLR4 (Sokolove et al., 2011).

The pathogenic influence of IgG autoantibody-ICs can be opposed by IgM natural antibodies to ACs. *In vivo* studies have shown that administration of anti-PC natural IgM greatly attenuated disease severity in a murine model of collagen-induced arthritis (CIA; Chen et al., 2009a). In this model, immunization with xenogenic collagen II emulsified in complete Freund's adjuvant induces a pathogenic autoimmune response to type II

collagen (Terato et al., 1992; Nandakumar et al., 2003), with tissue injury in part mediated through the activating Fc γ receptors (Kleinau et al., 2000). Infusions of IgM–NABs to ACs also blocked the disease process induced by passive transfer of anti-type II collagen autoantibodies (Chen et al., 2009a), in which inflammatory arthritis is mediated by Fc γ R and innate immune cells, while lymphocytes do not play central roles (Terato et al., 1992; Nandakumar et al., 2003).

In vitro studies have shown that anti-AC IgM antibodies can directly block the activating effects of lupus-associated IgG autoantibodies on bone marrow-derived DCs (Vas et al., 2012). In fact, the inflammatory effects of both anti-DNA and -RNA IgG-nucleic acid ICs in myeloid DCs were inhibited with suppression of the secretion of inflammatory cytokines IL-6 and TNF- α (Vas et al., 2012). This IgM–NAB also suppressed IC-mediated induction of cell surface expression of CD80 and CD86, as well as CD40 and other co-stimulatory molecules.

NATURAL ANTIBODY REGULATORY PATHWAYS THAT MODULATE INFLAMMATORY AND AUTOIMMUNE DISEASES

Serologic surveys of a large cohort of well-characterized SLE patients have further evaluated the potential clinical relevance of IgM autoantibodies to defined oxidation-associated antigenic-specificities, including the apoptosis-associated neo-antigens, PC and MDA. In the lupus cohort, levels of both of these types of IgM autoantibodies were significantly higher compared to healthy adult controls (Grönwall et al., 2012a). Importantly, higher levels of IgM anti-PC correlated with less long-term organ damage, as defined by the SLICC/ACR damage index score, as well as lower disease activity as assessed by the SELENA revision of the SLE disease activity index (SLEDAI) at the time of visit. IgM anti-PC levels also correlated with an absence of cardiovascular events, while there were no associations with renal disease (Grönwall et al., 2012a). These results are consistent with a previous report from a smaller cohort of Swedish patients (Su et al., 2008) and with studies showing that lower IgM anti-PC levels are associated with more frequent cardiovascular events in non-autoimmune patients (de Faire et al., 2010; Fiskesund et al., 2010). These findings were in fact

predicted by earlier studies in atherosclerosis-prone mice (Shaw et al., 2000). Indeed, pneumococcal vaccination, which induces PC-specific antibody responses, was shown to arrest plaque progression in LDL receptor-deficient mice with cholesterol levels over 1000 mg/dl (Binder et al., 2003). These findings have therefore further strengthened the hypothesis that some anti-AC IgM–Nabs can play protective roles in inflammatory disease.

Yet not every IgM–NAb that recognizes ACs may have equivalent clinical implications. In fact, levels of antibodies to PC and to MDA showed significant differences in their associations with lupus clinical manifestations. IgM anti-MDA showed only weak inverse correlations with the SLICC/ACR damage index but not the SELENA-SLEDAI score and there were also no significant associations with renal disease or cardiovascular events (Grönwall et al., 2012a). These studies also showed that higher levels of the IgM antibody to β 2-GPI correlated with less organ damage by SLICC/ACR damage index. Furthermore, patients without renal disease had higher levels of IgM anti-CL, and IgM anti-dsDNA (Grönwall et al., 2012a). This may indicate that higher levels of some IgM antibodies may protect some patients from kidney disease, as suggested in an earlier and more focused report (Mehrani and Petri, 2011). Taken together, these studies refute the notion that circulating IgM autoantibodies are inherently polyreactive. Instead, these data strongly argue that the antigenic fine binding specificity of the IgM determines whether there is an association with protection from certain lupus disease features (Grönwall et al., 2012a). In a recent clinical study, Ajeganova et al. (2011) examined RA patients treated with TNF- α blockers. They observed that levels of PC-specific natural IgM levels were increased in patients treated with TNF- α blockade, while lower anti-PC IgM levels correlated with inferior response to therapeutic intervention for RA disease. Further investigations will be needed to better understand how anti-AC Nabs may modulate the pathogenesis of different autoimmune rheumatic diseases.

MAPK PHOSPHATASE-1 IS REQUIRED FOR NATURAL ANTIBODY SUPPRESSION OF TLR RESPONSES

Investigations of signal transduction pathways have shown that IgM–Nabs to ACs can affect responses induced by agonists for a broad range of TLRs, by inhibition of the mitogen-activated protein kinase (MAPK) signal transduction system, which plays central roles in the induction and resolution of inflammatory responses (Grönwall et al., 2012b). Inflammatory responses can result from the induction of phosphorylation of one or more of the primary MAPKs; ERK1/2, JNK, and particularly p38, which then translocate to the nucleus where it can affect transcriptional regulation. In rheumatoid arthritis, activated (phosphorylated) p38 is increased in the RA synovium. However, despite evidence that small molecule p38 inhibitors have been effective in mouse models of inflammatory arthritis, efficacy in humans has been limited, which has suggested that an alternate approach to MAPK inhibition may provide greater clinical benefits (reviewed in Hammaker and Firestein, 2010).

To assess the potential relevance of this type of immunomodulatory NAb to clinical autoimmune diseases, the activity of the PC-specific IgM–Nab was also tested in a system in which inflammatory responses are induced by lupus IgG autoantibodies (Vas

et al., 2012). These studies demonstrated that this natural IgM inhibited p38 phosphorylation induced in DCs by nucleic acid-containing IgG autoantibody ICs (Vas et al., 2012). Notably, this inhibitory pathway also blocked the nuclear accumulation of the activated primary MAPKs in myeloid DCs (Vas et al., 2012). *In vitro* studies of murine bone marrow-derived DCs confirmed that this inhibition was entirely dependent on the recruitment of either C1q or MBL (Grönwall et al., 2012b; illustrated in Figure 1).

The magnitude and duration of MAPK signaling is dependent on the balance between the upstream activators of the system and the deactivation of these kinases by specific phosphatases. Based on evidence that this NAb could affect the activation of each of the three primary MAPKs (Grönwall et al., 2012b), studies were therefore performed that assessed the potential involvement of the regulatory MAPK phosphatases (MKPs), also known as dual-specificity phosphatases (DUSPs). These studies highlighted the role of MKP-1, the archetype for the family, which can serve as the counter-regulatory factor for all three of the primary MAPKs (reviewed in Liu et al., 2007). In fact, the anti-AC IgM-mediated blockade of TLR-mediated MAPK signaling had an absolute requirement for the expression of MKP-1 (Grönwall et al., 2012b). In DCs activated by TLR agonists, the addition of the anti-AC IgM, in the presence of C1q or MBL in serum-free media, resulted in induction within minutes of MKP-1 at a transcript and a protein level, and it rapidly became localized within the nucleus (Grönwall et al., 2012b). Using deconvolutional immunofluorescence microscopy, NAb-mediated MKP-1 accumulation correlated with a reciprocal impairment in the phosphorylation and nuclear translocation of the activated primary MAPK protein molecules. To investigate the relative contribution of MKP-1 to NAb-mediated suppression, responses were compared in DCs from wild-type or MKP-1-deficient mice (Grönwall et al., 2012b). Such MKP-1-deficient mice are reported to exhibit overexuberant inflammatory responses, but no other immune developmental abnormalities (Dorfman et al., 1996; Salojin et al., 2006). These studies confirmed the absolute requirement for MKP-1 for IgM–NAb-mediated inhibition of TLR responses from DCs (Grönwall et al., 2012b).

CONCLUDING REMARKS

One of the most fundamental challenges faced by the immune system is the efficient recognition and clearance of the body's own cells, which because of senescence or injury enter programmed cell death pathways. While cells dying of apoptotic death pathways do not pose an immediate risk to the host, if these cell corpses are not efficiently removed there is the risk of progression to secondary necrosis. This can lead to the loss of integrity of cell membranes with release of cytoplasmic and nuclear components that can serve as ligands for proinflammatory cellular receptors, and the triggering of autoimmune responses. Hence, throughout the lifespan of multicellular organisms, there is an absolute need for the clearance of the immense number of cell corpses that are generated each day, even in health. As a direct consequence, the immune system has developed a redundant layering of superimposed mechanisms. Hence, the control of apoptotic clearance is intertwined with the regulation and resolution of inflammatory responses.

At birth, humans already have substantial levels of circulating IgM antibodies, which reflect a functional B cell compartment poised and ready to contribute to neonatal host defenses. These IgM antibodies arise in the womb from neonatal B lymphocytes that express clonally distributed BCRs. However, evidence of recurrent clones suggests that this early B cell repertoire may be affected by *in vivo* clonal selection that may be a response to evolutionary pressure to provide important housekeeping functions related to apoptotic clearance and avoidance of excessive and damaging inflammatory responses.

IgM-mediated protection from autoimmune disease was first demonstrated in mice deficient in the capacity to secrete IgM antibodies, as these mice were found to develop pathologic autoimmunity with the production of lupus IgG autoantibodies (Boes et al., 2000; Ehrenstein et al., 2000). Furthermore, in mice predisposed to the development of lupus-like disease, a bias toward secretion of monomeric IgM and lower levels of polymeric IgM can result in accelerated development of lupus-like disease (Youd et al., 2004). It may therefore be relevant that 8% of a cohort of 300 SLE patients were recently reported to have selective deficiency in serum IgM (Perrazio et al., 2012).

Our studies demonstrated that anti-AC IgM-NAbs, present from early in life, can suppress inflammatory responses mediated by phagocytic cells by induction of MKP-1, which in other settings have been shown to have potent regulatory roles for the MAPK system. MKP-1 is well known for its many counter-regulatory roles, which include the late negative feedback of responses to LPS stimulation, the blunting of responses after rapid re-exposure to a TLR agonist such as LPS tolerance, as well as contributing to the anti-inflammatory properties of glucocorticoids (reviewed in Liu et al., 2007). These studies also documented an additive effect of anti-AC IgM-NAb and dexamethasone for early MKP-1 induction and inhibition of LPS-induced p38 MAPK activation that, when combined, exceeded the maximum effects of either agent alone (Grönwall et al., 2012b). In part, this is likely explained by the additive integration of separate signals received via distinct cell membrane-associated receptors triggered by dexamethasone (i.e., glucocorticoid receptor) or by anti-AC IgM complexes (discussed below). As glucocorticoids are amongst the most widely prescribed treatments for inflammatory and autoimmune diseases, it is indeed intriguing that there is an overlap in the inhibitory signal transduction pathways of glucocorticoids and by the formation of regulatory ICs with early complement recognition factors that are coordinated in their organization by IgM autoantibodies to oxidation-associated neo-determinants on ACs.

REFERENCES

- Ajegano, S., Fiskesund, R., de Faire, U., Hafström, I., and Frostegård, J. (2011). Effect of biological therapy on levels of atheroprotective antibodies against phosphorylcholine and apolipoproteins in rheumatoid arthritis – a one year study. *Clin. Exp. Rheumatol.* 29, 942–950.
- Alugupalli, K. R., and Abraham, D. (2009). B cell multitasking is required to control nematode infection. *Immunity* 230, 317–319.
- Arnold, J. N., Dwek, R. A., Rudd, P. M., and Sim, R. B. (2006). Mannan binding lectin and its interaction with immunoglobulins in health and in disease. *Immunol. Lett.* 106, 103–110.
- Asai, T., Tena, G., Plotnikova, J., Willmann, M. R., Chiu, W. L., Gomez-Gomez, L., et al. (2002). MAP kinase signalling cascade in *Arabidopsis* innate immunity. *Nature* 415, 977–983.
- Bagavant, H., and Fu, S. M. (2009). Pathogenesis of kidney disease in systemic lupus erythematosus. *Curr. Opin. Rheumatol.* 21, 489–494.
- Batista, B. S., Eng, W. S., Pilobello, K. T., Hendricks-Muñoz, K. D., and Mahal, L. K. (2011). Identification of a conserved glycan signature for microvesicles. *J. Proteome Res.* 10, 4624–4633.
- Baumgarth, N. (2011). The double life of a B-1 cell: self-reactivity selects for protective effector functions. *Nat. Rev. Immunol.* 11, 34–46.
- Binder, C. J., Horkko, S., Dewan, A., Chang, M.-K., Kieu, E. P., Goodyear, C. S., et al. (2003). Pneumococcal vaccination decreases atherosclerotic lesion formation: molecular mimicry between *Streptococcus pneumoniae* and oxidized LDL. *Nat. Med.* 9, 736–743.
- Boes, M., Schmidt, T., Linkemann, K., Beaudette, B. C., Marshak-Rothstein,

Fundamental to the inhibitory effects of regulatory NAbs, polymeric IgMs that bind ACs can express constant regions with multiple sites for recruitment of C1q, and the Fc μ of some IgM-NAbs also have high mannose glycoconjugates on that bind MBL (Chen et al., 2009a,b). The potential properties of such complexes suggested by studies with targeted deficiencies in C1q, MBL, or secreted IgM, which each have impaired control of inflammatory responses, and in some cases are predisposed to the development of autoimmune disease (Botto et al., 1998; Boes et al., 2000; Ehrenstein et al., 2000; Stuart et al., 2005). Furthermore, we have previously shown that the complement-dependent immunomodulatory properties of anti-AC IgM, while the recruitment of C1q or MBL was essential, there was no absolute requirement for downstream activation of the complement cascade (Chen et al., 2009a).

These IgM-NAbs to AC-associated determinants can regulate responses mediated by diverse TLRs, an ancient type of innate immune receptor that was first characterized in insects (Lemaitre et al., 1996). Furthermore, mechanistic investigations have shown these effects are linked to modulation of the MAPK signaling system, which is one of the earliest evolutionarily conserved pathways of immunity, being present in plants and mammals (Asai et al., 2002). Likewise, MKP-1 orthologs have also been described in protozoans (Moncho-Amor et al., 2011), and as mentioned above, mice with MKP-1 deficiency have severe defects in the control of innate responses (Salojin et al., 2006). These regulatory properties are expressed by a class of naturally occurring autoreactive antibodies that are postulated to come from the most primitive tier of B cells in the adaptive immune system (Kantor and Herzenberg, 1993).

As ACs are ubiquitous, we wonder whether the high frequency of these innate-like NAb-producing clones in the “preimmune” repertoire in part reflects positive selection of the B-1 cell clones that are reactive with membrane-associated neo-determinants on cells wasted during development. The protective properties of anti-AC NAb may be mediated by a previously unknown regulatory signaling pathway, which integrates and coordinates the influence of select innate immune factors on myeloid cell function.

ACKNOWLEDGMENTS

Work in our lab was supported by grants from the NIH; R01AI090118, R01 AI068063 and ARRA supplement, R01AI090118, and from the ACR REF Within Our Reach campaign, the Alliance for Lupus Research, the Arthritis Foundation, and the P. Robert Majumder Charitable Trust.

- A., and Chen, J. (2000). Accelerated development of IgG autoantibodies and autoimmune disease in the absence of secreted IgM. *Proc. Natl. Acad. Sci. U.S.A.* 97, 1184–1189.
- Botto, M., Dell'Agnola, C., Bygrave, A. E., Thompson, E. M., Cook, H. T., Petry, F., et al. (1998). Homozygous C1q deficiency causes glomerulonephritis associated with multiple apoptotic bodies. *Nat. Genet.* 19, 56–59.
- Boulé, M. W., Broughton, C., Mackay, F., Akira, S., Marshak-Rothstein, A., and Rifkin, I. R. (2004). Toll-like receptor 9-dependent and -independent dendritic cell activation by chromatin-immunoglobulin G complexes. *J. Exp. Med.* 199, 1631–1640.
- Cappione, A. III, Anolik, J. H., Pugh-Bernard, A., Barnard, J., Dutcher, P., Silverman, G. J., et al. (2005). Germinal center exclusion of autoreactive B cells is defective in human systemic lupus erythematosus. *J. Clin. Invest.* 115, 3205–3216.
- Casali, P., and Notkins, A. L. (1989). Probing the human B-cell repertoire with EBV: polyreactive antibodies and CD5⁺ lymphocytes. *Annu. Rev. Immunol.* 7, 513–535.
- Chen, Y., Khanna, S., Goodyear, C. S., Park, Y. B., Raz, E., Thiel, S., et al. (2009a). Regulation of dendritic cells and macrophages by an anti-apoptotic cell natural antibody that suppresses TLR responses and inhibits inflammatory arthritis. *J. Immunol.* 183, 1346–1359.
- Chen, Y., Park, Y. B., Patel, E., and Silverman, G. J. (2009b). IgM antibodies to apoptosis-associated determinants recruit C1q and enhance dendritic cell phagocytosis of apoptotic cells. *J. Immunol.* 182, 6031–6043.
- Chou, M. Y., Fogelstrand, L., Hartvigsen, K., Hansen, L. F., Woelkers, D., Shaw, P. X., et al. (2009). Oxidation-specific epitopes are dominant targets of innate natural antibodies in mice and humans. *J. Clin. Invest.* 119, 1335–1349.
- Cong, Y. Z., Rabin, E., and Wortis, H. H. (1991). Treatment of murine CD5[−] B cells with anti-Ig, but not LPS, induces surface CD5: two activation pathways. *Int. Immunol.* 3, 467–476.
- Cox, K. O., and Hardy, S. J. (1985). Autoantibodies against mouse bromelain-modified RBC are specifically inhibited by a common membrane phospholipid, phosphatidylcholine. *Immunology* 55, 263–269.
- Czajkowsky, D. M., and Shao, Z. (2009). The human IgM pentamer is a mushroom-shaped molecule with a flexural bias. *Proc. Natl. Acad. Sci. U.S.A.* 106, 14960–14965.
- Davies, D. R., Padlan, E. A., and Segal, D. M. (1975). Three-dimensional structure of immunoglobulins. *Annu. Rev. Biochem.* 44, 639–667.
- de Faire, U., Su, J., Hua, X., Frostegård, A., Halldin, M., Hellenius, M.-L., et al. (2010). Low levels of IgM antibodies to phosphorylcholine predict cardiovascular disease in 60-year old men: effects on uptake of oxidized LDL in macrophages as a potential mechanism. *J. Autoimmun.* 34, 73–79.
- Dorfman, K., Carrasco, D., Gruda, M., Ryan, C., Lira, S. A., and Bravo, R. (1996). Disruption of the erp/mkp-1 gene does not affect mouse development: normal MAP kinase activity in ERP/MKP-1-deficient fibroblasts. *Oncogene* 13, 925–931.
- Du Pasquier, L., Robert, J., Courtet, M., and Mußmann, R. (2000). B-cell development in the amphibian *Xenopus*. *Immunol. Rev.* 175, 201–213.
- Ehrenstein, M. R., Cook, H. T., and Neuberger, M. S. (2000). Deficiency in serum immunoglobulin (Ig)M predisposes to development of IgG autoantibodies. *J. Exp. Med.* 19, 1253–1258.
- Elkon, K., and Casali, P. (2008). Nature and functions of autoantibodies. *Nat. Clin. Pract. Rheumatol.* 4, 491–498.
- Feeney, A. J. (1991). Predominance of the T15 anti-phosphorylcholine junctional sequence in neonatal pre-B cell. *J. Immunol.* 147, 4343–4350.
- Feeney, A. J. (1992). Predominance of VH-D-JH junctions occurring at sites of short sequence homology results in limited junctional diversity in neonatal antibodies. *J. Immunol.* 149, 222–229.
- Feizi, T. (1988). Carbohydrate structures as onco-developmental antigens and components of receptor systems. *Adv. Exp. Med. Biol.* 228, 317–329.
- Ferry, H., Potter, P. K., Crockford, T. L., Nijnik, A., Ehrenstein, M. R., Walport, M. J., et al. (2007). Increased positive selection of B1 cells and reduced B cell tolerance to intracellular antigens in c1q-deficient mice. *J. Immunol.* 178, 2916–2922.
- Fiskesund, R., Stegmayer, B., Hallmans, G., Vikström, M., Weinshall, L., de Faire, U., et al. (2010). Low levels of antibodies against phosphorylcholine predict development of stroke in a population-based study from Northern Sweden. *Stroke* 41, 607–612.
- Flajnik, M., and Rumfelt, L. (2000). Early and natural antibodies in non-mammalian vertebrates. *Curr. Top. Microbiol. Immunol.* 252, 233–240.
- Fraser, D. A., Laust, A. K., Nelson, E. L., and Tenner, A. J. (2009). C1q differentially modulates phagocytosis and cytokine responses during ingestion of apoptotic cells by human monocytes, macrophages, and dendritic cells. *J. Immunol.* 183, 6175–6185.
- Friedman, P., Horkko, S., Steinberg, D., Witztum, J. L., and Dennis, E. A. (2002). Correlation of antiphospholipid antibody recognition with the structure of synthetic oxidized phospholipids. Importance of Schiff base formation and aldol condensation. *J. Biol. Chem.* 277, 7010–7020.
- Griffin, D. O., Holodick, N. E., and Rothstein, T. L. (2011). Human B1 cells in umbilical cord and adult peripheral blood express the novel phenotype CD20⁺CD27⁺CD43⁺CD70[−]. *J. Exp. Med.* 208, 67–80.
- Grillot-Courvalin, C., Brouet, J.-C., Labaume, S., Piller, F., Rassenti, L. Z., Silverman, G. J., et al. (1992). An anti-B cell autoantibody from Wiskott-Aldrich syndrome which recognizes i blood group specificity on normal human B cells. *Eur. J. Immunol.* 22, 1781–1788.
- Grönwall, C., Akhter, E., Oh, C., Burlingame, R. W., Petri, M., and Silverman, G. J. (2012a). IgM autoantibodies to distinct apoptosis-associated antigens correlate with protection from cardiovascular events and renal disease in patients with SLE. *Clin. Immunol.* 142, 390–398.
- Grönwall, C., Chen, Y., Vas, J., Khanna, S., Thiel, S., Corr, M., et al. (2012b). MAPK phosphatase-1 is required for regulatory natural autoantibody-mediated inhibition of TLR responses. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19745–19750.
- Hammaker, D., and Firestein, G. S. (2010). “Go upstream, young man”: lessons learned from the p38 saga. *Ann. Rheum. Dis.* 69, i77–i82.
- Hardy, R. R. (2006). B-1 B cell development. *J. Immunol.* 177, 2749–2754.
- Hardy, R. R., and Hayakawa, K. (2005). Development of B cells producing natural autoantibodies to thymocytes and senescent erythrocytes. *Springer Semin. Immunopathol.* 26, 363–375.
- Hayakawa, K., Asano, M., Shinton, S. A., Gui, M., Allman, D., Stewart, C. L., et al. (1999). Positive selection of natural autoreactive B cells. *Science* 285, 113–116.
- Jackson, K. J., Wang, Y., Gaeta, B. A., Pomat, W., Siba, P., Rimmer, J., et al. (2012). Divergent human populations show extensive shared Igk rearrangements in peripheral blood B cells. *Immunogenetics* 64, 3–14.
- Jenne, C. N., Kennedy, L. J., and Reynolds, J. D. (2006). Antibody repertoire development in the sheep. *Dev. Comp. Immunol.* 30, 165–174.
- Kantor, A. B., and Herzenberg, L. A. (1993). Origin of murine B lineages. *Annu. Rev. Immunol.* 11, 501–538.
- Kearney, J. F. (2005). Innate-like B cells. *Springer Semin. Immunopathol.* 26, 377–383.
- Kearney, J. F., Barletta, R., Quan, Z. S., and Quintáns, J. (1981). Monoclonal vs. heterogeneous anti-H-8 antibodies in the analysis of the anti-phosphorylcholine response in BALB/c mice. *Eur. J. Immunol.* 11, 877–883.
- Kleinau, S., Martinsson, P., and Heyman, B. (2000). Induction and suppression of collagen-induced arthritis is dependent on distinct fcgamma receptors. *J. Exp. Med.* 191, 1611–1616.
- Kobie, J. J., Alcena, D. C., Zheng, B., Bryk, P., Mattiacio, J. L., Brewer, M., et al. (2012). 9G4 autoreactivity is increased in HIV-infected patients and correlates with HIV broadly neutralizing serum activity. *PLoS ONE* 7:e35356. doi: 10.1371/journal.pone.0035356
- Korb, L. C., and Ahearn, J. M. (1997). C1q binds directly and specifically to surface blebs of apoptotic human keratinocytes: complement deficiency and systemic lupus erythematosus revisited. *J. Immunol.* 158, 4525–4528.
- Kulik, L., Fleming, S. D., Moratz, C., Reuter, J. W., Novikov, A., Chen, K., et al. (2009). Pathogenic natural antibodies recognizing annexin IV are required to develop intestinal ischemia-reperfusion injury. *J. Immunol.* 182, 5363–5373.
- Leadbetter, E. A., Rifkin, I. R., Hohlbach, A. M., Beaudette, B. C., Shlomchik, M. J., and Marshak-Rothstein, A. (2002). Chromatin-IgG complexes activate B cells by dual engagement of IgM and Toll-like receptors. *Nature* 416, 603–607.
- Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J. M., and Hoffmann, J. A. (1996). The dorsoventral regulatory gene cassette spätzle/Toll/cactus controls the potent antifungal response in *Drosophila* adults. *Cell* 86, 973–983.
- Liu, Y., Shepherd, E. G., and Nelis, L. D. (2007). MAPK phosphatases: regulating the immune response. *Nat. Rev. Immunol.* 7, 202–212.

- Lleo, A., Invernizzi, P., Gao, B., Podda, M., and Gershwin, M. E. (2010). Definition of human autoimmunity – autoantibodies versus autoimmune disease. *Autoimmun. Rev.* 9, A259–A266.
- Lövgren, T., Eloranta, M.-L., Båvè, U., Alm, G. V., and Rönnblom, L. (2004). Induction of interferon- α production in plasmacytoid dendritic cells by immune complexes containing nucleic acid released by necrotic or late apoptotic cells and lupus IgG. *Arthritis Rheum.* 50, 1861–1872.
- Lutz, H. U. (ed.) (2012). *Naturally Occurring Antibodies (Nabs)*. Austin, TX: Lands Bioscience.
- Masmoudi, H., Mota-Santos, T., Huett, F., Coutinho, A., and Cazenave, P. A. (1990). All T15 Id-positive antibodies (but not of VHT15⁺ antibodies) are produced by peritoneal CD5⁺ B lymphocytes. *Int. Immunol.* 2, 515–520.
- Matsushita, M., Matsushita, A., Endo, Y., Nakata, M., Kojima, N., Mizuochi, T., et al. (2004). Origin of the classical complement pathway: lamprey orthologue of mammalian C1q acts as a lectin. *Proc. Natl. Acad. Sci. U.S.A.* 101, 10127–10131.
- Mehrani, T., and Petri, M. (2011). IgM anti- β 2 glycoprotein I is protective against lupus nephritis and renal damage in systemic lupus erythematosus. *J. Rheumatol.* 38, 450–453.
- Mercolino, T. J., Arnold, L. W., and Haughton, G. (1986). Phosphatidyl choline is recognized by a series of Ly-1⁺ murine B cell lymphomas specific for erythrocyte membranes. *J. Exp. Med.* 163, 155–165.
- Mi, Q. S., Zhou, L., Schulze, D. H., Fischer, R. T., Lustig, A., Rezanka, L. J., et al. (2000). Highly reduced protection against *Streptococcus pneumoniae* after deletion of a single heavy chain gene in mouse. *Proc. Natl. Acad. Sci. U.S.A.* 97, 6031–6036.
- Moncho-Amor, V., Galardi-Castilla, M., Perona, R., and Sastre, L. (2011). The dual-specificity protein phosphatase MkpB, homologous to mammalian MKP phosphatases, is required for *D. discoideum* post-aggregative development and cisplatin response. *Differentiation* 81, 199–207.
- Nandakumar, K. S., Svensson, L., and Holmdahl, R. (2003). Collagen type II-specific monoclonal antibody-induced arthritis in mice: description of the disease and the influence of age, sex, and genes. *Am. J. Pathol.* 163, 1827–1837.
- Navratil, J. S., Watkins, S. C., Wisnieski, J. J., and Ahearn, J. M. (2001). The globular heads of C1q specifically recognize surface blebs of apoptotic vascular endothelial cells. *J. Immunol.* 166, 3231–3239.
- Ogden, C. A., Decathelineau, A., Hoffmann, P. R., Bratton, D., Ghebrehiwet, B., Fadok, V. A., et al. (2001). C1q and mannose binding lectin engagement of cell surface calreticulin and CD91 initiates macropinocytosis and uptake of apoptotic cells. *J. Exp. Med.* 194, 781–795.
- Paidassi, H., Tacnet-Delorme, P., Garlatti, V., Darnault, C., Ghebrehiwet, B., Gaboriaud, C., et al. (2008). C1q binds phosphatidylserine and likely acts as a multiligand-bridging molecule in apoptotic cell recognition. *J. Immunol.* 180, 2329–2338.
- Pascual, V., and Capra, J. D. (1992). VH4-21, a human VH gene segment overrepresented in the autoimmune repertoire. *Arthritis Rheum.* 35, 11–18.
- Pascual, V., Victor, K., Lelsz, D., Spellerberg, M., Hamblin, T., Thompson, K., et al. (1991). Nucleotide sequence analysis of the V regions of two IgM cold agglutinins. Evidence that the VH4-21 gene segment is responsible for the major cross-reactive idiotyp. *J. Immunol.* 146, 4385–4391.
- Perlmuter, R., Kearney, J., Chang, S., and Hood, L. (1985). Developmentally controlled expression of immunoglobulin VH genes. *Science* 227, 1597–1601.
- Perrazio, S. F., Salomao, R., Silva, N. P., Carneiro-Sampaio, M., and Andrade, L. E. C. (2012). Serial screening shows that 28% of systemic lupus erythematosus adult patients carry an underlying primary immunodeficiency. *Arthritis Rheum.* 64, S284.
- Pugh-Bernard, A. E., Silverman, G. J., Cappione, A. J., Villano, M. E., Ryan, D. H., Insel, R. A., et al. (2001). Regulation of inherently autoreactive VH4-34 B cells in the maintenance of human B cell tolerance. *J. Clin. Invest.* 108, 1061–1070.
- Quartier, P., Potter, P. K., Ehrenstein, M. R., Walport, M. J., and Botto, M. (2005). Predominant role of IgM-dependent activation of the classical pathway in the clearance of dying cells by murine bone marrow-derived macrophages *in vitro*. *Eur. J. Immunol.* 35, 252–260.
- Rogosch, T., Kerzel, S., Hoß, K., Hoersch, G., Zemlin, C., Heckmann, M., et al. (2012). IgA response in preterm neonates shows little evidence of antigen-driven selection. *J. Immunol.* 189, 5449–5456.
- Salojin, K. V., Owusu, I. B., Millerchip, K. A., Potter, M., Platt, K. A., and Oravecz, T. (2006). Essential role of MAPK phosphatase-1 in the negative control of innate immune responses. *J. Immunol.* 176, 1899–1907.
- Schroeder, H., Hillson, J., and Perlmuter, R. (1987). Early restriction of the human antibody repertoire. *Science* 238, 791–793.
- Schroeder, H. W., and Wang, J. Y. (1990). Preferential utilization of conserved immunoglobulin heavy chain variable gene segments during human fetal life. *Proc. Natl. Acad. Sci. U.S.A.* 87, 6146–6150.
- Seidl, K. J., MacKenzie, J. D., Wang, D., Kantor, A. B., Kabat, E. A., Herzenberg, L. A., et al. (1997). Frequent occurrence of dentic heavy, and light chain Ig rearrangements. *Int. Immunol.* 9, 689–670.
- Shaw, P. X., Höök, S., Chang, M.-K., Curtiss, L. K., Palinski, W., Silverman, G. J., et al. (2000). Natural antibodies with the T15 idiomotype may act in atherosclerosis, apoptotic clearance, and protective immunity. *J. Clin. Invest.* 105, 1731–1740.
- Shaw, P. X., Goodey, C. S., Chang, M. K., Witztum, J. L., and Silverman, G. J. (2003). The autoreactivity of anti-phosphorylcholine antibodies for atherosclerosis-associated neo-antigens and apoptotic cells. *J. Immunol.* 170, 6151–6157.
- Silberstein, L., Jefferies, L., Goldman, J., Friedman, D., Moore, J., Nowell, P., et al. (1991). Variable region gene analysis of pathologic human autoantibodies to the related i and I red blood cell antigens. *Blood* 78, 2372–2386.
- Silverman, G. J., Chen, P., and Carson, D. (1990). Cold agglutinins: specificity, idiotypic and structural analysis. *Chem. Immunol.* 48, 109–125.
- Sokolove, J., Zhao, X., Chandra, P. E., and Robinson, W. H. (2011). Immune complexes containing citrullinated fibrinogen costimulate macrophages via Toll-like receptor 4 and Fcg receptor. *Arthritis Rheum.* 63, 53–62.
- Stuart, L. M., Takahashi, K., Shi, L., Savill, J., and Ezekowitz, R. A. (2005). Mannose-binding lectin-deficient mice display defective apoptotic cell clearance but no autoimmune phenotype. *J. Immunol.* 174, 3220–3226.
- Su, J., Hua, X., Concha, H., Svennungsson, E., Cederholm, A., and Frostegård, J. (2008). Natural antibodies against phosphorylcholine as potential protective factors in SLE. *Rheumatology* 47, 1144–1150.
- Sun, J., Hayward, C., Shinde, R., Christenson, R., Ford, S. P., and Butler, J. E. (1998). Antibody repertoire development in fetal and neonatal piglets. I. Four VH genes account for 80 percent of VH usage during 84 days of fetal life. *J. Immunol.* 161, 5070–5078.
- Terato, K., Hasty, K., Reife, R., Cremer, M., Kang, A., and Stuart, J. (1992). Induction of arthritis with monoclonal antibodies to collagen. *J. Immunol.* 148, 2103–2108.
- Vas, J., Grönwall, C., Marshak-Rothstein, A., and Silverman, G. J. (2012). Natural antibody to apoptotic cell membranes inhibits the proinflammatory properties of lupus autoantibody immune complexes. *Arthritis Rheum.* 64, 3388–3398.
- Wardemann, H., Yurasov, S., Schaefer, A., Young, J. W., Meffre, E., and Nussenzweig, M. C. (2003). Predominant autoantibody production by early human B cell precursors. *Science* 301, 1374–1377.
- Yasuda, K., Richez, C., Maciaszek, J. W., Agrawal, N., Akira, S., Marshak-Rothstein, A., et al. (2007). Murine dendritic cell type I IFN production induced by human IgG-RNA immune complexes is IFN regulatory factor (IRF)5 and IRF7 dependent and is required for IL-6 production. *J. Immunol.* 178, 6876–6885.
- Youd, M. E., Luus, L., and Corley, R. B. (2004). IgM monomers accelerate disease manifestations in autoimmune-prone Fas-deficient mice. *J. Autoimmun.* 23, 333–343.
- Zhang, M., Alicot, E. M., and Carroll, M. C. (2008). Human natural IgM can induce ischemia/reperfusion injury in a murine intestinal model. *Mol. Immunol.* 245, 1036–1039.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 August 2012; accepted: 03 January 2013; published online: 05 February 2013.

Citation: Vas J, Grönwall C and Silverman GJ (2013) Fundamental roles of the innate-like repertoire of natural antibodies in immune homeostasis. Front. Immunol. 4:4. doi: 10.3389/fimmu.2013.00004

This article was submitted to Frontiers in B Cell Biology, a specialty of Frontiers in Immunology.

Copyright © 2013 Vas, Grönwall and Silverman. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



Differences in the composition of the human antibody repertoire by B cell subsets in the blood

Eva Szymanska Mroczek¹, Gregory C. Ippolito², Tobias Rogosch³, Kam Hon Hoi^{4,5}, Tracy A. Hwangpo⁶, Marsha G. Brand⁶, Yingxin Zhuang⁶, Cun Ren Liu⁶, David A. Schneider⁷, Michael Zemlin³, Elizabeth E. Brown⁸, George Georgiou^{2,4,5} and Harry W. Schroeder Jr.^{1,6*}

¹ Department of Microbiology, University of Alabama at Birmingham, Birmingham, AL, USA

² Department of Molecular Biosciences, University of Texas at Austin, Austin, TX, USA

³ Laboratory for Neonatology and Pediatric Immunology, Department of Pediatrics, Philipps-University, Marburg, Germany

⁴ Department of Chemical Engineering, University of Texas at Austin, Austin, TX, USA

⁵ Department of Biomedical Engineering, University of Texas at Austin, Austin, TX, USA

⁶ Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA

⁷ Department of Biochemistry and Molecular Genetics, University of Alabama at Birmingham, Birmingham, AL, USA

⁸ Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA

Edited by:

Ignacio Sanz, University of Rochester, USA

Reviewed by:

I-Hsin Su, Nanyang Technological University, Singapore

Masaki Hikida, Kyoto University, Japan

***Correspondence:**

Harry W. Schroeder Jr., Department of Medicine, University of Alabama at Birmingham, SHEL 176, 1530 3rd Avenue South, Birmingham, AL 35294-2182, USA
e-mail: hwsj@uab.edu

The vast initial diversity of the antibody repertoire is generated centrally by means of a complex series of V(D)J gene rearrangement events, variation in the site of gene segment joining, and TdT catalyzed N-region addition. Although the diversity is great, close inspection has revealed distinct and unique characteristics in the antibody repertoires expressed by different B cell developmental subsets. In order to illustrate our approach to repertoire analysis, we present an in-depth comparison of V(D)J gene usage, hydrophobicity, length, D_H reading frame, and amino acid usage between heavy chain repertoires expressed by immature, transitional, mature, memory IgD⁺, memory IgD⁻, and plasmacytes isolated from the blood of a single individual. Our results support the view that in both human and mouse, the H chain repertoires expressed by individual, developmental B cell subsets appear to differ in sequence content. Sequencing of unsorted B cells from the blood is thus likely to yield an incomplete or compressed view of what is actually happening in the immune response of the individual. Our findings support the view that studies designed to correlate repertoire expression with diseases of immune function will likely require deep sequencing of B cells sorted by subset.

Keywords: human antibody repertoire, CDR-H3, B cells subsets

INTRODUCTION

Production of a highly diverse, polyclonal immunoglobulin repertoire plays a central role in the ability of B cells to produce antibodies specific to a diverse range of foreign and self-antigens (1, 2). The antigen-binding sites of these antibodies are created by the juxtaposition of six hypervariable loops, termed complementarity determining regions (CDRs): three from the heavy (H) and three from the light (L) chain V domains. Because the third CDR of the H chain, termed CDR-H3 (2–5), is the direct product of V(D)J joining and N-region addition, it is the most variable component of the pre-immune immunoglobulin repertoire. The location of CDR-H3 at the center of the antigen-binding site allows this interval to play a key role in antigen recognition and binding (6–8).

Developing B cells pass through a series of checkpoints designed to test the functionality and antigen specificity of the immunoglobulin (9–14). In adults, this process begins in the bone marrow, and then continues in the periphery where it is heavily influenced by exposure to both self and foreign antigens. Immature B cells are released into the blood and in the periphery pass through a transitional stage prior to entering specific anatomic sites, such as the splenic marginal zone and the splenic and lymph node follicles (15, 16). Maturation is associated with the

co-expression of IgM and IgD (17). Mature cells exposed to antigen can become either memory cells or plasmacytes. Both types of cells circulate through the blood on their way to their specific anatomic niches (18–21). IgM bearing memory cells can be divided into two populations, those that express IgD concurrently and those that do not (22–25). The IgM⁺IgD⁻ memory B cell population includes conventional, follicular B cells, whereas the IgM⁺IgD⁺ memory B cell population includes marginal zone-like B cells that play a more immediate role in response to foreign antigens (26–28).

Recent studies in mice have shown that the composition of CDR-H3 exhibits preferred patterns in amino acid composition, length, and charge distribution that differ by developmental stage and B cell subset (29–33). These categorical constraints are initially imposed by natural selection of the germline V, D, and J gene sequence; and alteration of the sequence of these gene segments can give rise to dramatically different CDR-H3 repertoires (34–36). D gene sequence-specific changes in CDR-H3 content lead to altered patterns of B cell development, antigen-specific antibody production, and levels of protection against infectious agents (31, 37, 38), which underscores the important role played by the composition of the CDR-H3 repertoire in the regulation and function of the humoral immune response.

Given the importance of CDR-H3 to antigen recognition and antibody specificity, and the observation that CDR-H3 content can differ by peripheral developmental stage in the mouse; we sought to test whether V(D)J usage and CDR-H3 content would also differ by developmental stage in human. We used surface expression of CD19, CD27, IgD, CD24, and CD38 expression to identify and sort immature, transitional, mature, memory IgD⁺, memory IgD⁻ B cell subsets, and plasmacytes from the blood of a healthy female subject. We then used RT-PCR followed by Roche GS-FLX 454 deep sequencing to clone and sequence Cμ and Cγ-containing transcripts from the sorted cells. As in the mouse, we found that the distribution of V, D, and J utilization, and CDR-H3 length, amino acid usage, and average hydrophobicity differed between developmentally and functionally distinct B cell subsets. We conclude that studies of differences between healthy individuals and patients with diseases referable to the humoral immune response will likely require comparisons of the B cell repertoire by subset.

MATERIALS AND METHODS

SUBJECT DESCRIPTION AND ISOLATION OF B CELL SUBSETS

One healthy female subject, age 56, was recruited for antibody repertoire high throughput sequencing using the 454 platform. The subject is Caucasian, a lifelong native of the state of Alabama, and was without a history of illness or repeated infection that could be related to abnormal immune function. The complete blood count was well within normal limits. Serum immunoglobulin levels were IgM 382, IgG 1,680, and IgA 368 mg/dL, respectively. Venous blood (100 cm³) was drawn by routine venipuncture and mononuclear cells were isolated using Ficoll-Paque Plus (GE Healthcare). CD19⁺ magnetic beads (Miltenyi Biotec MACS) were used to enrich for B cells. These CD19⁺ cells were further fractionated by CD27[±] populations using CD27 magnetic beads (Miltenyi Biotec MACS) according to the manufacturer's protocol. CD19⁺CD27⁺ B cells were stained with CD19 APC₇₈₀ (eBioscience), CD27 PE-Cy7 (BD Pharmingen), CD24 APC (BioLegend), and IgD FITC (Southern Biotech), and sorted into IgD⁺ memory B cells (CD19⁺/CD27⁺/IgD⁺/CD24⁺), IgD⁻ memory B cells (CD19⁺/CD27⁺/IgD⁻/CD24⁺), and plasmacytes (CD19⁺/CD27⁺/CD24⁻) using a high speed sorting cytometer (FACSAria III; Becton Dickinson). CD19⁺CD27⁻ B cells were stained with CD19 APC₇₈₀ (eBioscience), CD24 APC (BioLegend), CD38 PE (BioLegend), and IgD FITC (Southern Biotech) and sorted into mature/naïve (CD19⁺/CD27⁻/IgD⁺/CD38⁺/CD24⁺), transitional (CD19⁺/CD27⁻/IgD⁺/CD38⁺⁺⁺/CD24⁺⁺⁺), and immature (CD19⁺/CD27⁻/IgD⁻) B cell subsets. Each B cell subset was then individually resuspended in 1 mL TRI reagent (Ambion) and archived at -80°C until processed for total RNA extraction. This work was performed in accordance with an Institutional Review Board approved protocol and informed consent was obtained from the subject at the University of Alabama at Birmingham, Birmingham, AL, USA.

GENERATION OF IgH LIBRARIES

For RNA extraction, 0.2 mL chloroform was added to the 1 mL sample, vortexed for 15 s, left to stand at room temperature for 5 min, then spun at 12,000 × g for 10 min at 4°C. The aqueous

phase (~400 μL) was removed and to this an equal volume of 70% ethanol was added and then mixed by pipetting. This was applied immediately to an RNA-binding silica spin-column and subsequently processed according to the manufacturer's protocol (Qiagen RNeasy micro column; catalog no. 74004). Purified total RNA was eluted in 14 μL RNase-free water. Oligo-dT primer was used to generate first-strand cDNA from ~100 ng input RNA using the SuperScript RT II synthesis kit (Invitrogen; catalog no. 11904-018) per the manufacturer's protocol.

FastStart high fidelity PCR system (Roche; catalog no. 03-553-361-001) and an equimolar mix of eight optimized VH-FWD primers previously described for human IgH amplification (39, 40) coupled with a multiplex of 10-nucleotide uniquely barcoded CH-REV primers: IgM-rev, 5'-10 nt ID-GGTTGGGGCGGATGCCTCC-3', and IgG-all-rev, 5'-10 nt ID-SGATGGGCCCTGGTGARGC-3' were used to amplify V(D)JCμ and V(D)JCγ cDNAs from the cDNA template. Cycling conditions were as follows: 95°C denaturation for 3 min; 92°C for 1 min, 50°C for 1 min, 72°C for 1 min for 4 cycles; 92°C for 1 min, 55°C for 1 min, 72°C for 1 min for 4 cycles; 92°C for 1 min, 63°C for 1 min, 72°C for 1 min for 22 cycles; 72°C for 7 min. PCR amplicons were gel-purified (Zymo Research) before sequencing.

HIGH-THROUGHPUT SEQUENCING OF IgH REPERTOIRES AND BIOINFORMATIC ANALYSIS

The University of Texas Genomics Sequencing and Analysis Facility performed Roche GS-FLX 454 deep sequencing. CH-REV barcodes were examined to verify the integrity of each library after filtering raw data for read quality. Sequences were submitted to the ImMunoGeneTics (IMGT) database and IMGT/high V-QUEST web-based analysis tool (version 1.0.3) (41). The 11 CSV text files outputted by IMGT/highV-QUEST were then imported into IgAT immunoglobulin analysis tool for further deconstruction (42). Differences between populations were assessed, where appropriate, by Student's *t*-test, two tailed; Fisher's exact test, two tailed and *d*; χ², or Levene's test for the homogeneity of variance. Analysis was performed with PRISM version 5 (Graph Pad). The standard deviation accompanies mean. Raw 454 sequence files were deposited to the NCBI Sequence Read Archive (Accession SRP037774).

RESULTS

ISOLATION OF B LINEAGE CELLS AND 454 HIGH-THROUGHPUT SEQUENCING OF IgH TRANSCRIPTS FROM PERIPHERAL BLOOD

CD19⁺ cells bearing the cell surface markers characteristic of immature, transitional, mature, memory IgD⁺, memory IgD⁻, and plasmacytes were isolated from the blood of a healthy female subject (43–47) (Figure 1). Following total RNA extraction, PCR was used to amplify cDNA copies of V(D)JCμ and V(D)JCγ transcripts using optimized VH-FWD primers previously described for human IgH amplification (39, 40). We obtained a total of 15,433 immature, 37,396 transitional, 47,781 mature, 43,558 memory IgD⁺, 28,142 memory IgD⁻, and 43,824 plasmacyte unique and in-frame IgH heavy chain reads. Of these, we obtained 1,240 immature, 1,354 transitional, 1,250 mature, 1,244 memory IgD⁺, 833 memory IgD⁻, and 1,714 plasmacyte reads that were of sufficient length to be identified as Igμ sequences, and 1,879

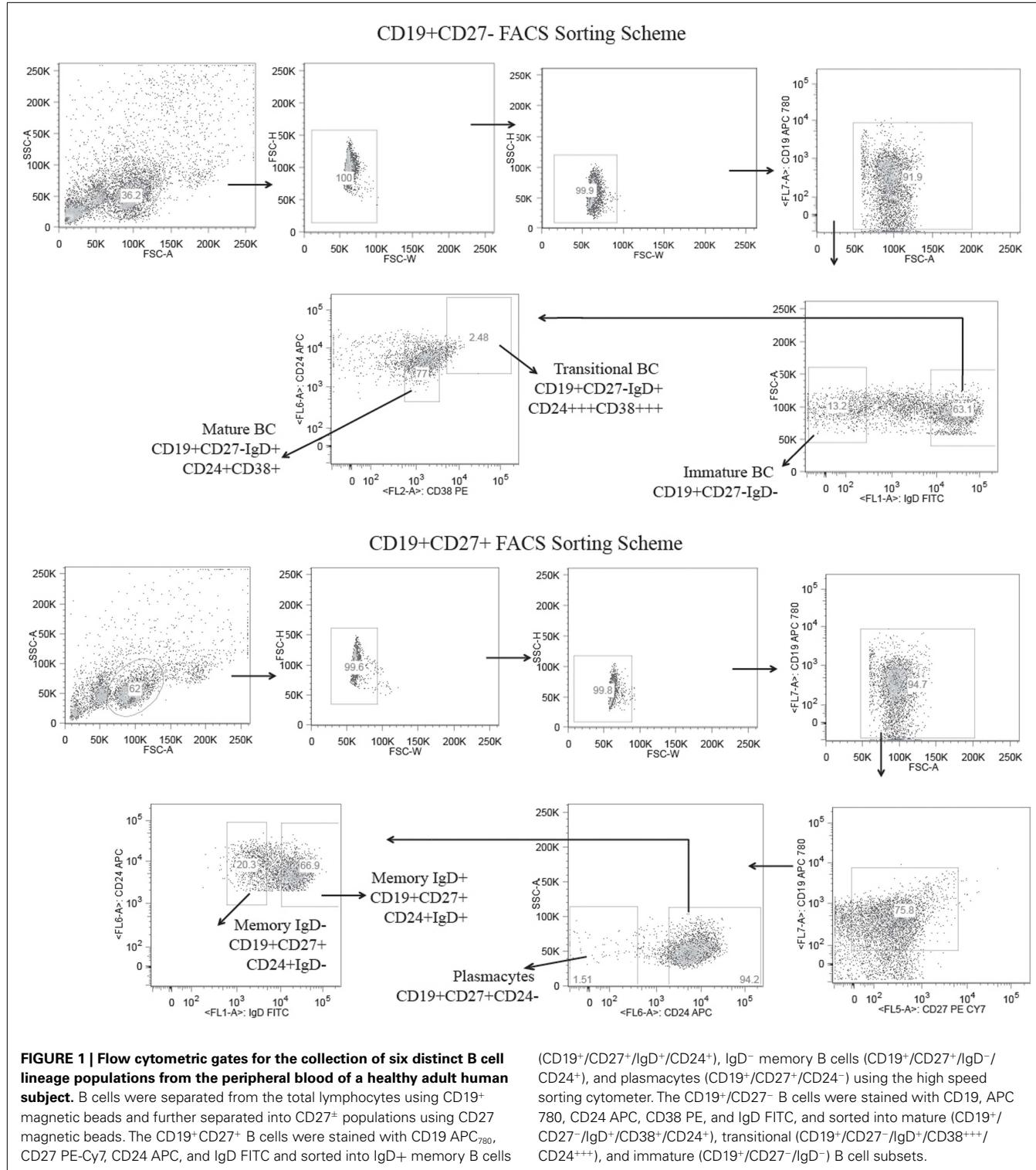


FIGURE 1 | Flow cytometric gates for the collection of six distinct B cell lineage populations from the peripheral blood of a healthy adult human subject. B cells were separated from the total lymphocytes using CD19+ magnetic beads and further separated into CD27 \pm populations using CD27 magnetic beads. The CD19+CD27+ B cells were stained with CD19 APC₇₈₀, CD27 PE-Cy7, CD24 APC, and IgD FITC and sorted into IgD+ memory B cells

(CD19 $^{+}$ /CD27 $^{+}$ /IgD $^{+}$ /CD24 $^{+}$), IgD $^{-}$ memory B cells (CD19 $^{+}$ /CD27 $^{+}$ /IgD $^{-}$ /CD24 $^{+}$), and plasmacytes (CD19 $^{+}$ /CD27 $^{+}$ /CD24 $^{-}$) using the high speed sorting cytometer. The CD19 $^{+}$ /CD27 $^{-}$ B cells were stained with CD19, APC₇₈₀, CD24 APC, CD38 PE, and IgD FITC, and sorted into mature (CD19 $^{+}$ /CD27 $^{-}$ /IgD $^{+}$ /CD38 $^{+++}$ /CD24 $^{+++}$), transitional (CD19 $^{+}$ /CD27 $^{-}$ /IgD $^{+}$ /CD38 $^{+++}$ /CD24 $^{+++}$), and immature (CD19 $^{+}$ /CD27 $^{-}$ /IgD $^{-}$) B cell subsets.

memory IgD $^{-}$ and 3,347 plasmacyte reads that were of sufficient length to be identified as Ig γ sequences. All of the unique Ig μ and Ig γ reads were deconstructed to assess the presence and extent of changes in these repertoires that had occurred as B cells progressed through the various developmental checkpoints.

THE IMMATURE B CELL RECEPTOR REPERTOIRE UTILIZES SHORTEST CONTRIBUTION OF GERMLINE GENE VJ SEGMENTS AND FAVORS V1–18, D2–15, D4–23, AND D5–12

The immature B cell subset is primarily composed of recent bone marrow emigrants. It expressed a highly diverse repertoire that

differed from the subsequent transitional stage in that it contained the smallest contribution of germline V and J gene sequence to the CDR-H3 region (**Figure 2**). By family, V_{H4} gene segments contributed the most, followed by V_{H3} , V_{H1} , V_{H5} , V_{H2} , and V_{H6} (**Figure 3**). By individual V gene segments, V1–18, V1–69, V3–73, and V4–59 were most common. Across subsets, the immature B cell

subset was enriched for V1–18, V3–30–3, and V3–74 (**Figure 4**). By D_H family, D_{H3} was the most common, followed by D_{H2} and D_{H6} (**Figure 5**). By individual D gene segment, D2–2, D3–3, D3–22, D6–13, and D6–19 were favored. Across subsets, D1–26, D2–15, D3–10, D4–23, and D5–12 were more commonly used in the immature B cell lineage (**Figure 6**). By J_H gene segment,

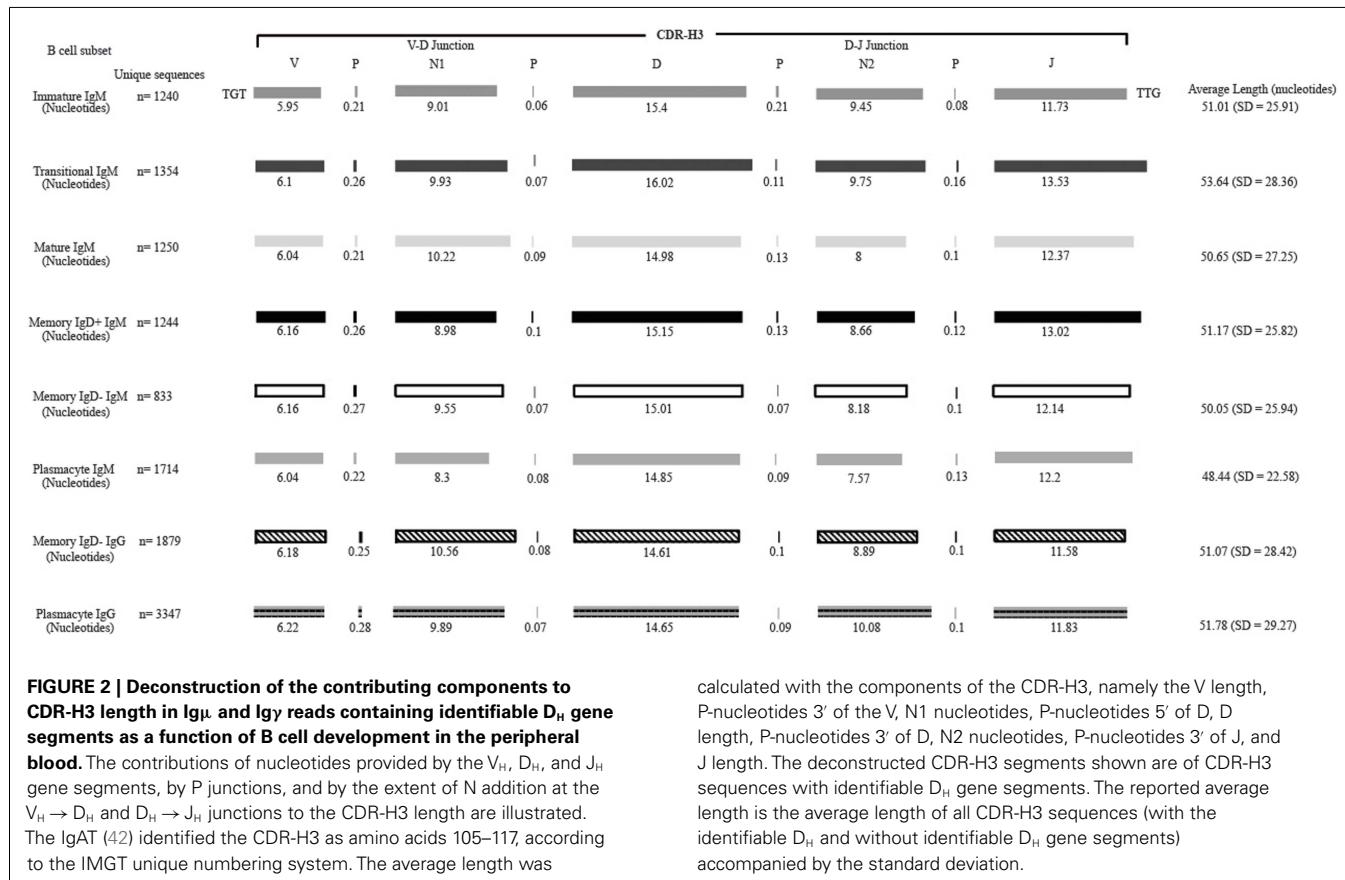


FIGURE 2 | Deconstruction of the contributing components to CDR-H3 length in IgM and IgY reads containing identifiable D_H gene segments as a function of B cell development in the peripheral blood. The contributions of nucleotides provided by the V_H , D_H , and J_H gene segments, by P junctions, and by the extent of N addition at the $V_H \rightarrow D_H$ and $D_H \rightarrow J_H$ junctions to the CDR-H3 length are illustrated. The IgAT (42) identified the CDR-H3 as amino acids 105–117, according to the IMGT unique numbering system. The average length was

calculated with the components of the CDR-H3, namely the V length, P-nucleotides 3' of the V, N1 nucleotides, P-nucleotides 5' of D, D length, P-nucleotides 3' of D, N2 nucleotides, P-nucleotides 3' of J, and J length. The deconstructed CDR-H3 segments shown are of CDR-H3 sequences with identifiable D_H gene segments. The reported average length is the average length of all CDR-H3 sequences (with the identifiable D_H and without identifiable D_H gene segments) accompanied by the standard deviation.

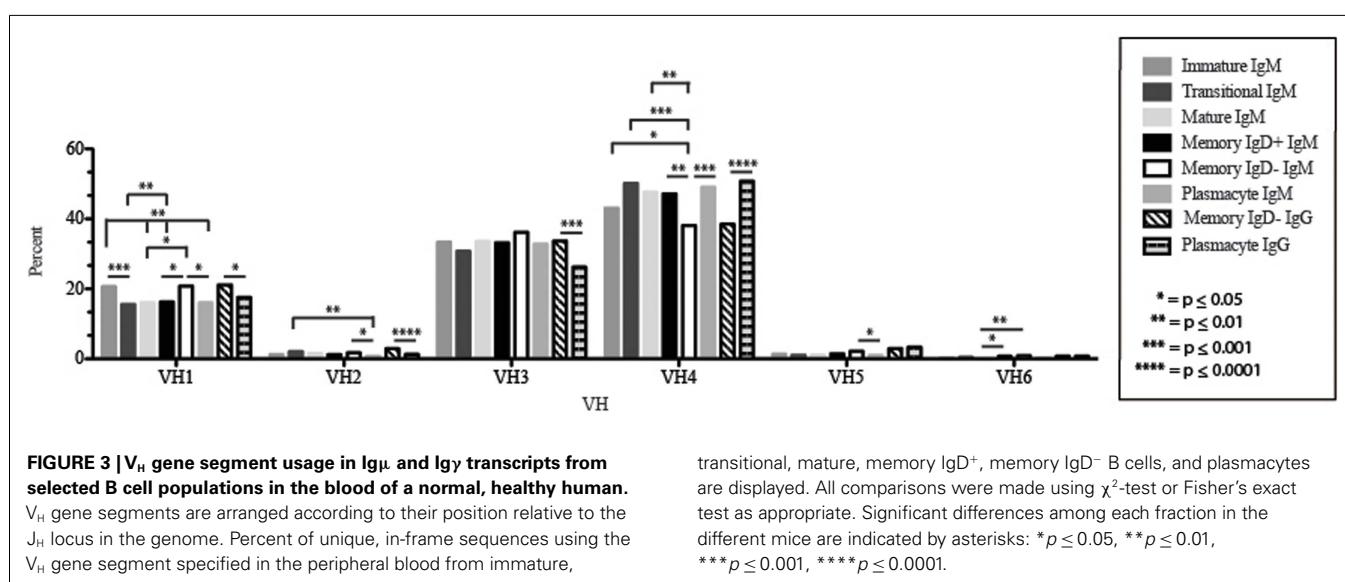
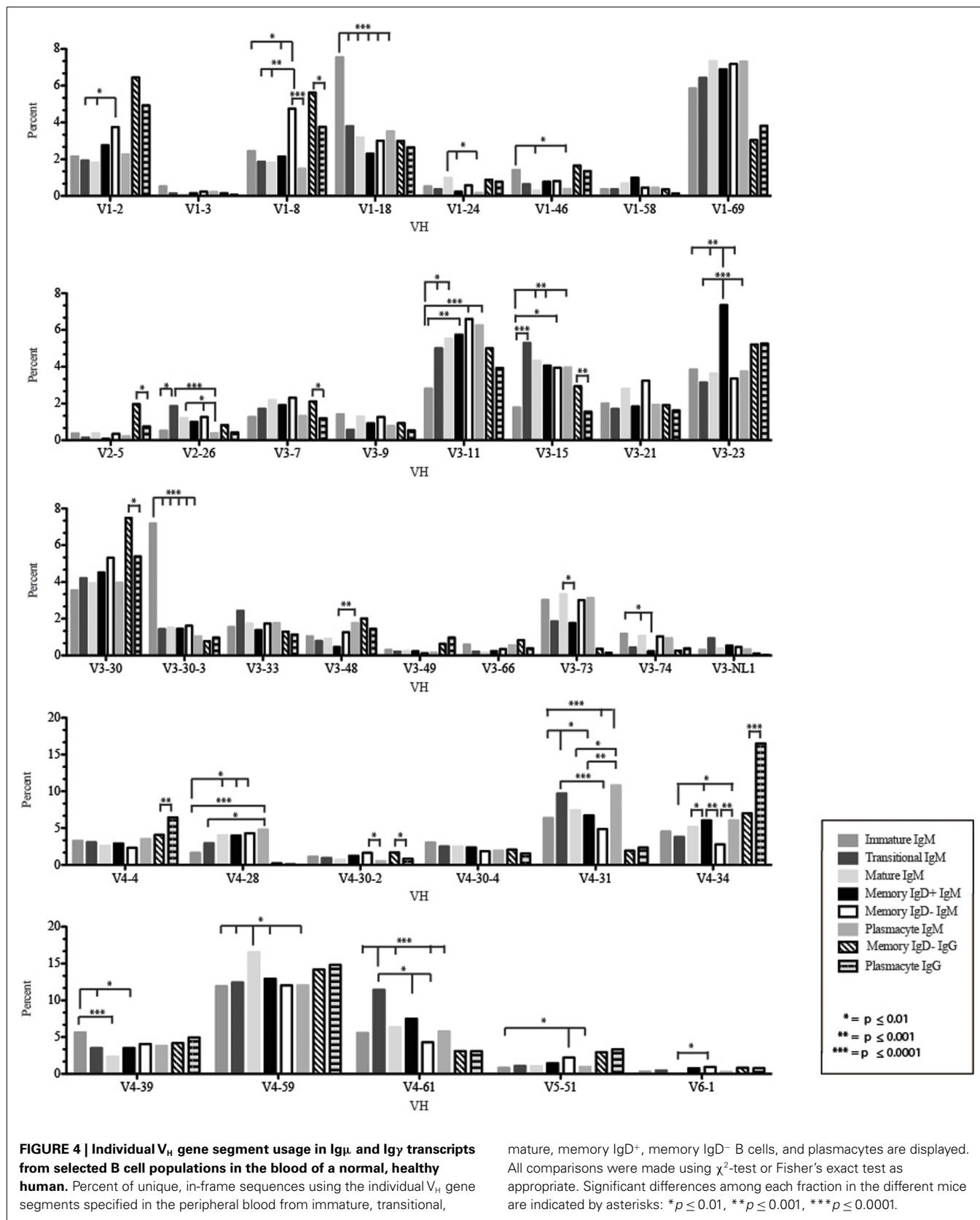
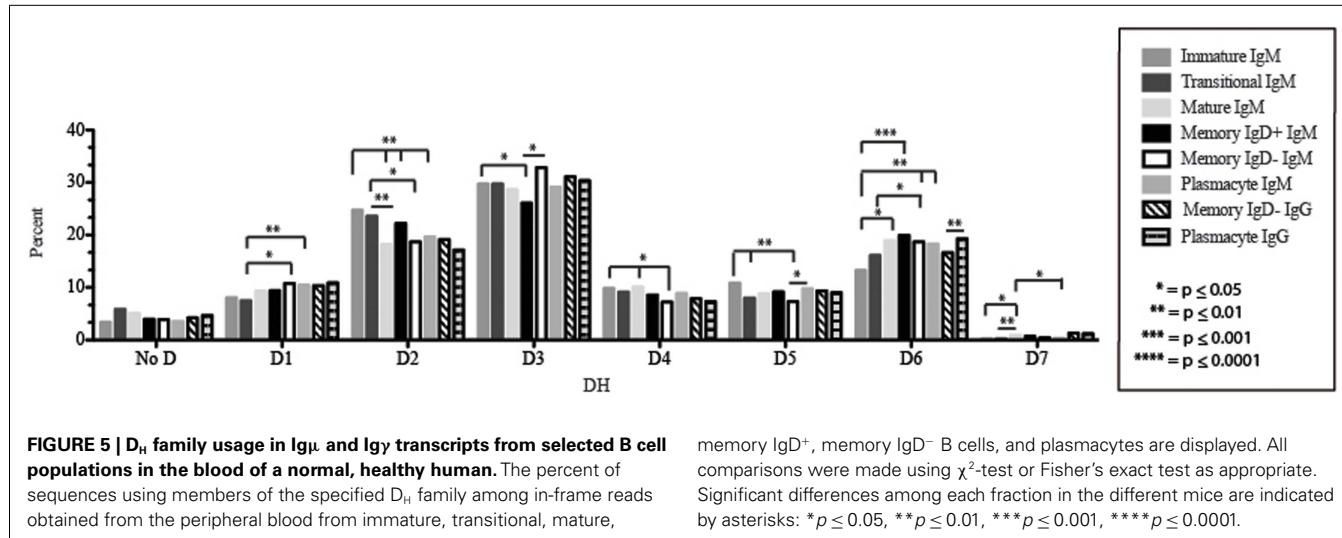


FIGURE 3 | V_H gene segment usage in IgM and IgY transcripts from selected B cell populations in the blood of a normal, healthy human. V_H gene segments are arranged according to their position relative to the J_H locus in the genome. Percent of unique, in-frame sequences using the V_H gene segment specified in the peripheral blood from immature, transitional, mature, memory IgD+, memory IgD-, and plasmacytes are displayed. All comparisons were made using χ^2 -test or Fisher's exact test as appropriate. Significant differences among each fraction in the different mice are indicated by asterisks: * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.

transitional, mature, memory IgD+, memory IgD-, B cells, and plasmacytes are displayed. All comparisons were made using χ^2 -test or Fisher's exact test as appropriate. Significant differences among each fraction in the different mice are indicated by asterisks: * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.





J_H4 was the most common, followed by J_H6, J_H5, and J_H3. Across subsets, immature B cells used J_H5 more frequently (**Figure 7**).

Amino acid usage in the CDR-H3 loops expressed by these immature B cells varied within a narrow range. When compared to transitional cells, immature B cells used less arginine, asparagine ($p = 0.02$), aspartic acid (0.04), glutamine ($p = 0.009$), glutamic acid ($p = 0.02$), tyrosine ($p = 0.002$), threonine ($p = 0.0039$), cysteine ($p < 0.0001$), and leucine ($p = 0.02$) (**Figure 8**). As a result of the decrease in the use of hydrophobic and hydrophilic amino acids, the immature repertoire exhibited the lowest prevalence of highly hydrophobic (hydrophobicity >0.7) CDR-H3 loops ($p < 0.05$) and the lowest prevalence of the highly hydrophilic (hydrophobicity ≤ 0.7) CDR-H3 loops of the six subsets examined (**Figure 9**).

THE TRANSITIONAL B CELL REPERTOIRE IS CHARACTERIZED BY THE LONGEST CDR-H3 LOOP LENGTH, INCREASED USE OF D2-2, AND INCREASED USE OF TYROSINE

Of the six subsets examined, the transitional CDR-H3 repertoire was the most heavily enriched for longest CDR-H3 loops (**Figure 2**). This bias for increased length reflects greater preservation of V(D)J gene segment sequence (**Figure 2**). Conversely, transitional B cell CDR-H3s were enriched for N nucleotide addition, averaging total 19.68 nucleotides and 9.75 nucleotides at the D → J junction (**Figure 2**). This was the first in a general pattern of diminishing N addition with maturation. Compared to the immature B cell fraction, there was a significant decrease for V_H1 family gene segments ($p < 0.001$) (**Figure 3**). By V gene segment, the use of V1-69, V2-26, V3-7, V3-11, V3-15, V3-21, V3-30, V3-33, V3-NL1, V4-28, V4-31, V4-61 was greater than in immature B cells, whereas use of V1-2, V1-3, V1-8, V1-18, V1-24, V1-46, V1-58, V2-5, V3-9, V3-21, V3-23, V3-30-3, V3-48, V3-66, V3-73, V3-74, V4-34, and V4-39 was decreased (**Figure 4**). The transitional B cell CDR-H3 loop utilized higher levels of D_H6 gene segments (not significant), with lower levels of D_H5 ($p = 0.005$) than immature B cells (**Figure 5**). By D gene assignment, a significant increase in D1-1 ($p = 0.09$) and D2-2 ($p = 0.0002$) usage in transitional B cells

memory IgD⁺, memory IgD⁻ B cells, and plasmacytes are displayed. All comparisons were made using χ^2 -test or Fisher's exact test as appropriate. Significant differences among each fraction in the different mice are indicated by asterisks: * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.

was observed when compared with the immature fraction, with a compensatory decrease in D1-26 ($p = 0.01$), D2-15 ($p < 0.0001$), D3-10 ($p = 0.005$), D4-23 ($p = 0.0026$), and D5-12 ($p = 0.0004$) (**Figure 6**). The use of J_H6 ($p = 0.0008$) was greater than in immature B cells, while the use of J_H4 ($p = 0.01$) and J_H5 ($p = 0.09$) was decreased (**Figure 7**).

CDR-H3 loops of these transitional cells used more arginine, lysine, asparagine ($p = 0.02$), aspartic acid ($p = 0.04$), glutamine ($p = 0.009$), glutamic acid ($p = 0.02$), tyrosine ($p = 0.001$), threonine ($p = 0.003$), cysteine ($p < 0.0001$), and leucine ($p = 0.02$), while using less tryptophan, serine, glycine, alanine, methionine, and phenylalanine than immature B cells (**Figure 8**). Of the six subsets studied, transitional B cells exhibited the higher prevalence of charged sequences as compared to the immature fraction (**Figure 9**). The contrast to the immature population was the most striking, suggesting specific gain of charged CDR-H3s in the transition from the immature to the transitional B cell stage. Conversely, the prevalence of highly hydrophobic CDR-H3s increased when compared to the immature B cell fraction.

THE MATURE B CELL SUBSET DEMONSTRATES A DECREASE IN THE USAGE OF DH2 AND JH6, AND AN INCREASE IN THE PERCENTAGE OF HIGHLY HYDROPHOBIC AND CHARGED CDR-H3 LOOPS

The mature B cell population was at the median for total CDR-H3 length and for the relative contributions of germline (**Figure 2**). Conversely, mature B cell CDR-H3s were enriched for N nucleotide addition, averaging 18.22 nucleotides total and 10.22 nucleotides at the V → D junction (**Figure 2**). In comparison to the transitional B cell repertoire, mature B cells exhibited similar expression of V_H family gene usage (**Figure 3**). An increase in V4-59 ($p = 0.01$) and a decrease in the use of V4-61 ($p < 0.0001$), respectively, were observed when compared to the transitional and mature fractions (**Figure 4**). Use of the D_H2 ($p = 0.01$) family in general, and the D2-2 gene segment ($p = 0.01$) in particular, was lower than in transitional cells (**Figures 5 and 6**). There was an increase in the use of J_H1 ($p = 0.0004$) with a decrease in the use of J_H6 ($p = 0.002$) (**Figure 7**).

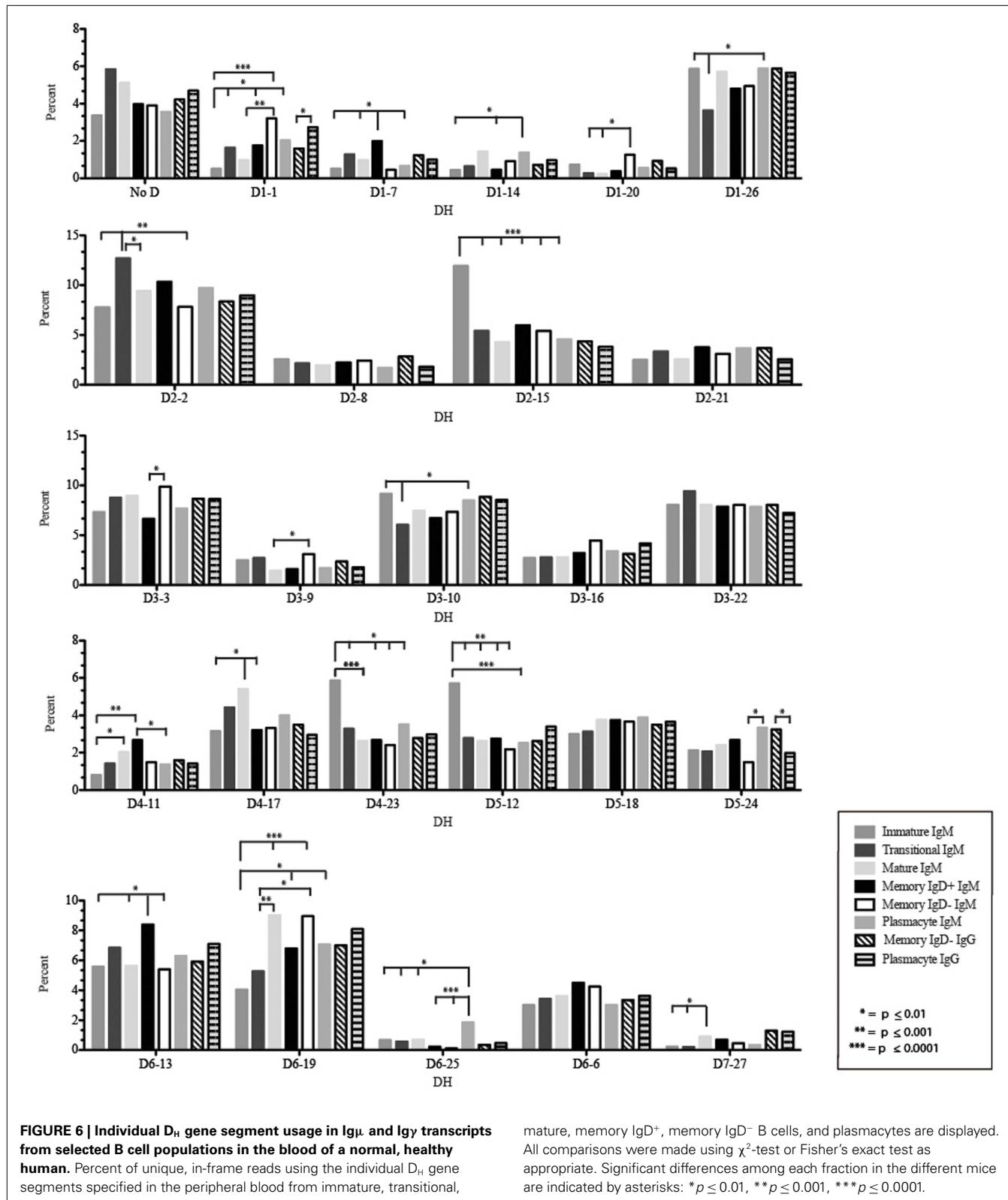
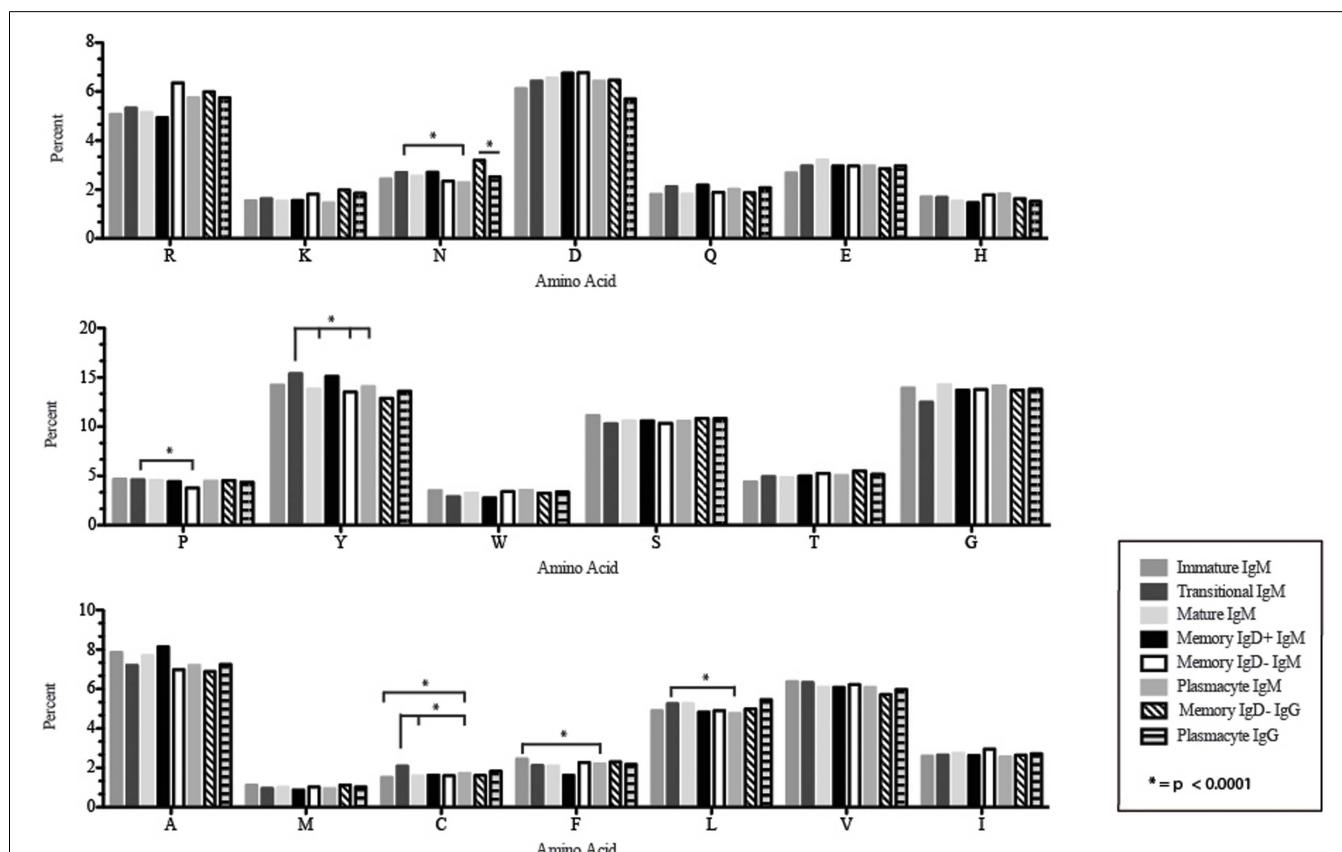
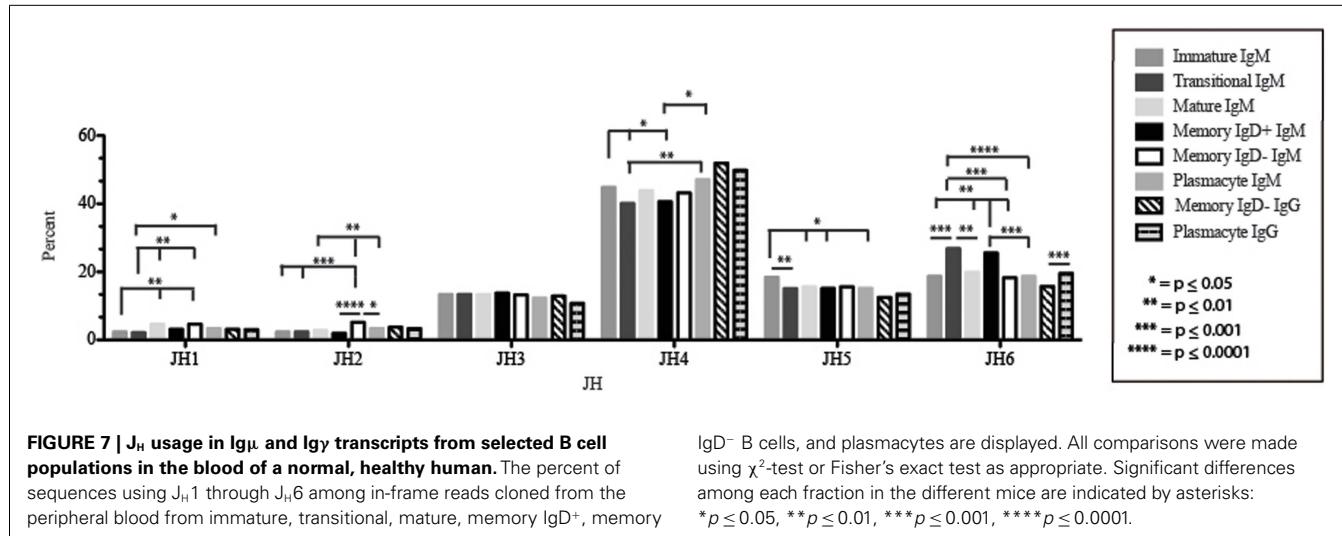


FIGURE 6 | Individual D_H gene segment usage in Igμ and Igγ transcripts from selected B cell populations in the blood of a normal, healthy human. Percent of unique, in-frame reads using the individual D_H gene segments specified in the peripheral blood from immature, transitional,

mature, memory IgD⁺, memory IgD⁻ B cells, and plasmacytes are displayed. All comparisons were made using χ^2 -test or Fisher's exact test as appropriate. Significant differences among each fraction in the different mice are indicated by asterisks: * $p \leq 0.01$, ** $p \leq 0.001$, *** $p \leq 0.0001$.

CDR-H3 loops demonstrated an increase in the use of glutamine ($p = 0.007$), with a decrease in tyrosine ($p = < 0.0001$), cysteine ($p = 0.0001$), and valine ($p = 0.04$) (Figure 8). As a result,

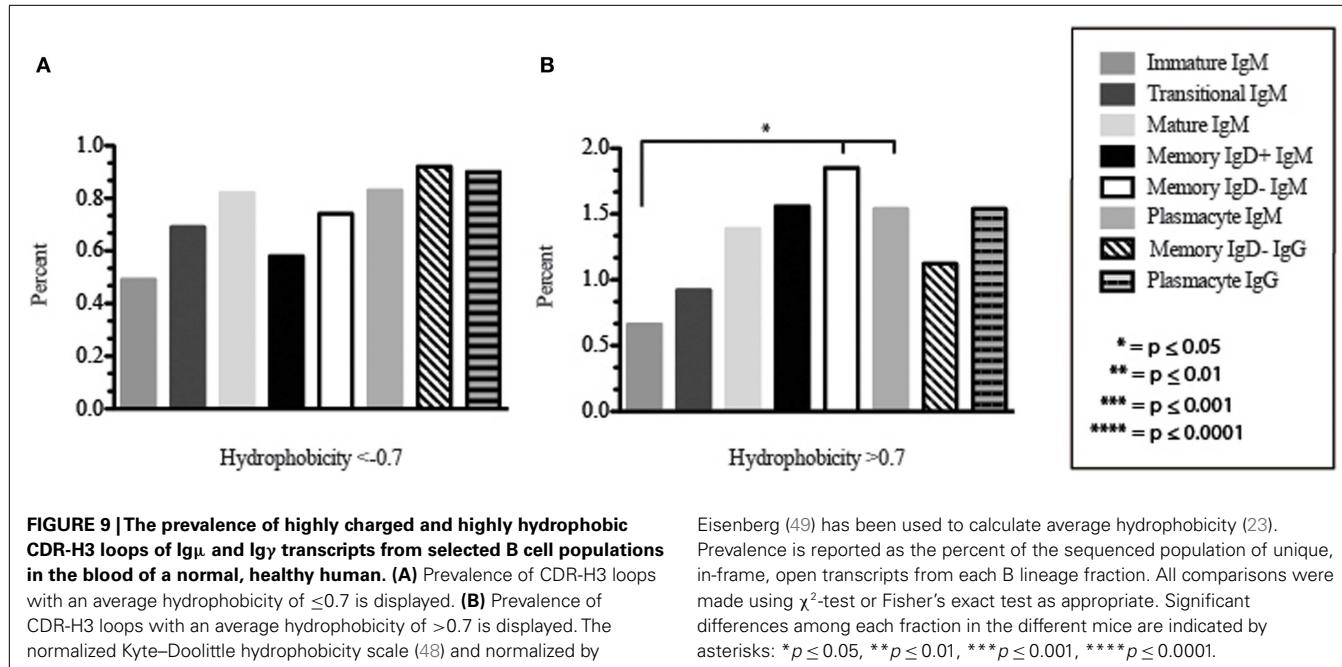
the mature B cell repertoire was enriched for the use of hydrophobic and charged CDR-H3 loops when compared with immature and transitional subsets (Figure 9).



MEMORY IgD⁺ AND IgD⁻ B CELLS DISPLAY DIVERGENT Ig_μ REPERTOIRES

The Ig_μ repertoires of the memory IgD⁺ and memory IgD⁻ blood B cells were distinguishable and divergent from both mature B cells

and from each other. The memory IgD⁺ B cell CDR-H3 region exhibited a greater contribution of germline D_H and J_H gene sequences than memory IgD⁻ (Figure 2). Memory IgD⁺ B cells used V_H4 ($p = 0.008$) family gene segments more frequently than



memory IgD[−] B cells, and V_H1 ($p = 0.03$) family gene segments less frequently. The memory IgD[−] B cells used V_H1 ($p = 0.03$) gene segments more frequently and V_H4 ($p = 0.03$) gene segments less frequently than mature B cells (Figure 3). By individual gene V_H gene segment, the most prominent differences between memory IgD⁺ and IgD[−] reflected increased use of V3–23 ($p = 0.0003$), V4–34 ($p = 0.001$), V4–61 ($p = 0.004$) in the former, and decreased use of V1–8 ($p = 0.002$) and V4–74 ($p = 0.02$) in the latter ($p < 0.0001$) (Figure 4), with the exception of V4–31 ($p = 0.02$, memory IgD[−]) and V4–59 ($p = 0.02$, memory IgD⁺ and $p = 0.01$, memory IgD[−]), which was increased among mature B cells (Figure 4).

Igμ from memory IgD⁺ B cells used D3 ($p = 0.01$) family D_H gene segments less frequently than memory IgD[−] cells (Figure 5). When compared with mature B cells, the memory IgD⁺ Igμ repertoire also used D2 and family D_H gene segments more frequently and D3 family D_H gene segments less frequently (not significant). Finally, memory IgD[−] B cells appeared to use D3 family D_H gene segments more frequently than mature B cells, although this preference did not achieve statistical significance. By individual D_H gene segment, the memory IgD⁺ Igμ repertoire displayed increased use of D6–13 ($p = 0.01$); and a decrease in use of D3–3 ($p = 0.01$) (Figure 6). Divergent usage of J_H2 and J_H6 was also observed (Figure 7). The memory IgD⁺ Igμ repertoire used J_H6 more frequently than the memory IgD[−] ($p = 0.001$) or mature B cell Igμ repertoire ($p = 0.009$); and J_H2 ($p < 0.0001$) less frequently than memory IgD[−]. J_H usage in the memory IgD[−] Igμ repertoire was very similar to that observed in mature B cells, with the exception of an increase in memory IgD[−] J_H2 usage as compared to the mature B cells ($p = 0.007$) (Figure 7).

The CDR-H3 loop of the memory IgD⁺ B Igμ repertoire contained more proline ($p = 0.01$), tyrosine ($p = 0.01$), and alanine ($p = 0.005$); but less arginine ($p = 0.001$), and tryptophan ($p = 0.04$) than memory IgD[−] B cells (Figure 8). The increase in

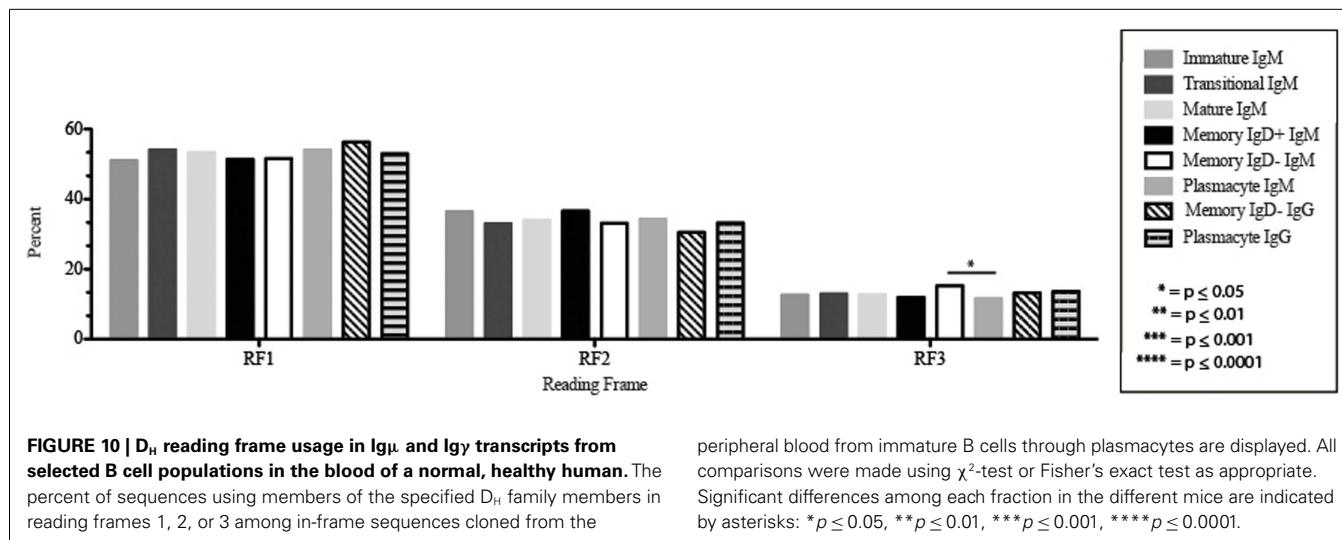
tyrosine reflected increased use of J_H6, rather than increased use of reading frame 1. Indeed, use of reading frame 1, 2, and 3 were similar between the memory fractions (Figure 10). When compared to mature B cells, the memory IgD⁺ Igμ repertoire was similarly enriched for glutamine ($p = 0.02$) and tyrosine ($p = 0.03$), and depleted of phenylalanine ($p = 0.01$). The memory IgD[−] Igμ repertoire also contained more arginine ($p = 0.005$) and less proline ($p = 0.01$) than the mature B cell Igμ repertoire. The memory IgD[−] Igμ repertoire relatively contained a higher percentage of highly charged CDR-H3s (hydrophobicity > 0.7) (1.85%) when compared to the Igμ repertoires of subsequent B cell fractions (Figure 9).

THE PLASMACYTE Igμ REPERTOIRE DIVERGED FROM BOTH THE MEMORY IgD⁺ AND IgD[−] Igμ REPERTOIRE, AS WELL AS FROM THE MATURE B CELL Igμ REPERTOIRE

In comparison to the other Igμ and Igγ repertoires, the CDR-H3 component of the plasmacyte Igμ repertoire exhibited the fewest N nucleotides at both the V → D and D → J junctions, respectively. As a result, not only the Igμ repertoire relatively enriched for germline V(D)J sequence, but also exhibited the shortest average length (Figure 2).

By V_H family, plasmacytes exhibited higher usage of V_H4 than either memory B cell population, and lower usage of V_H2, V_H3, and V_H5 (Figure 3). These differences were most affected by increased use of V4–34 ($p = 0.007$, $p < 0.0001$) when compared to both the memory IgD[−] and IgD⁺ Igμ repertoires and decreased use of V5–51 ($p = 0.01$) when compared to the memory IgD[−] Igμ repertoire (Figure 4).

The distribution of D_H gene family usage among the plasmacyte Igμ repertoire was similar to that of the mature B cell Igμ repertoire, but differed for individual families with the two memory B cell Igμ repertoires. There were no statistically significant



differences in the use of D_H gene segments between the memory IgD $^+$ and the plasmacyte Ig μ repertoires. When compared to the memory IgD $-$ Ig μ repertoire, the plasmacyte Ig μ repertoire used D_H 5 gene segments more frequently ($p = 0.04$) (Figure 5). By individual D_H gene segment, plasmacytes used D6–25 more frequently ($p = 0.006$) and D7–27 less frequently ($p = 0.03$) than mature B cells. Plasmacytes used D6–25 more frequently ($p < 0.0001$), and D4–11 ($p = 0.01$), D6–13 ($p = 0.04$), and D6–6 ($p = 0.03$) less frequently than the IgD $^+$ memory Ig μ repertoire. Finally, plasmacytes used D5–24 ($p = 0.007$) and D6–25 ($p = 0.0001$) more frequently, and D3–9 ($p = 0.03$) less frequently than the memory IgD $-$ Ig μ repertoire (Figure 6).

By J_H gene segment, the plasmacyte Ig μ repertoire displayed similar levels of J gene segments when compared to the mature B cell Ig μ repertoire. Plasmacytes expressed higher levels of J_H 2 ($p = 0.01$), J_H 4 ($p = 0.01$); and lower levels of J_H 6 than memory IgD $^+$ B cells ($p = 0.0004$). Finally, plasmacytes expressed lower levels of J_H 2 ($p = 0.04$) than memory IgD $-$ B cells (Figure 7).

When compared with the mature B cell Ig μ repertoire, plasmacytes expressed lower levels of asparagine ($p = 0.02$), alanine ($p = 0.01$), and leucine ($p = 0.007$) in the CDR-H3 loop. When compared with memory IgD $^+$ B cells, plasmacytes expressed lower levels of asparagine ($p = 0.001$), aspartic acid ($p = 0.02$), glutamine ($p = 0.04$), tyrosine ($p = 0.001$), and alanine ($p = 0.0002$); and higher levels of tryptophan ($p = 0.02$) and phenylalanine ($p = 0.01$). When compared with the memory IgD $-$ Ig μ repertoire, plasmacytes expressed lower levels of arginine ($p = 0.02$), lysine ($p = 0.009$), and isoleucine ($p = 0.02$) (Figure 8).

When comparing the relative prevalence of either highly charged or highly hydrophobic CDR-H3 loops, plasmacytes were enriched for charged CDR-H3 loops (0.84%) in comparison to the five other Ig μ repertoires (Figures 9 and 11). The distribution of highly hydrophobic CDR-H3 loops decreased in plasmacytes (1.54%) as compared to memory IgD $-$ B cells (1.85%), and returned to the comparable levels of memory IgD $^+$ B cells (1.56%) (Figure 9).

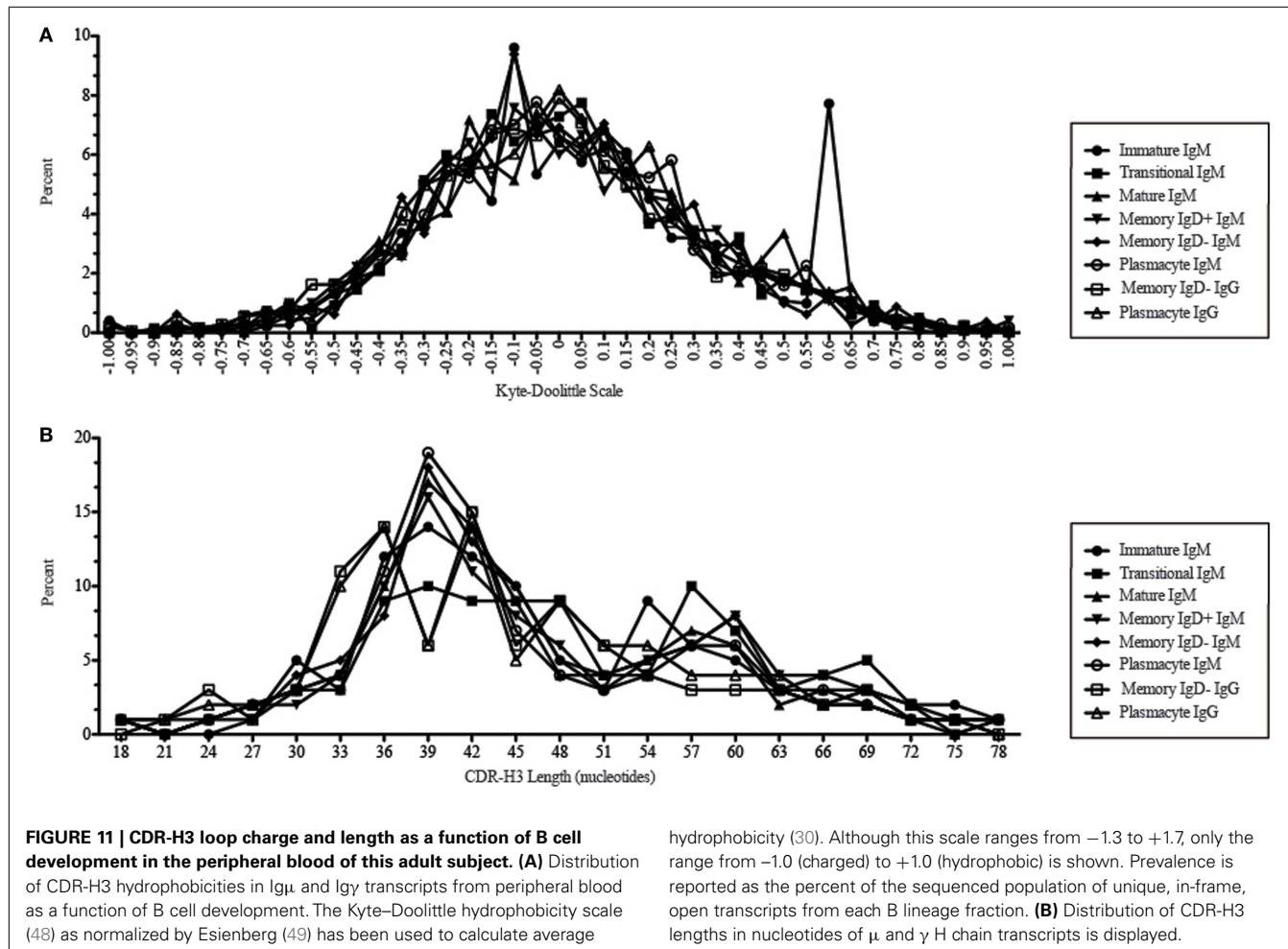
peripheral blood from immature B cells through plasmacytes are displayed. All comparisons were made using χ^2 -test or Fisher's exact test as appropriate. Significant differences among each fraction in the different mice are indicated by asterisks: * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.

THE PLASMACYTE Ig γ REPERTOIRE DIVERGED FROM IgD $-$ IgD $^+$ MEMORY B CELLS

The Ig γ repertoires expressed by memory IgD $-$ B cells and plasmacytes were distinguishable and uniquely different from each other. While the average length and V(D)J gene segment length was very similar between the memory IgD $-$ and plasmacytes, differences in the N-region additions were observed. The memory IgD $-$ B cell CDR-H3 region exhibited a greater number of N nucleotide addition at the V-D junction (10.56 nucleotides) as compared to the plasmacytes. Conversely, plasmacytes contained more N nucleotide addition at the D-J junction than memory IgD $-$ B cells (10.08 nucleotides) (Figure 2). Memory IgD $-$ B cells used V_H 1 ($p = 0.03$), V_H 2 ($p = 0.0001$), and V_H 3 ($p = 0.0003$) family gene segments more frequently than plasmacytes; and V_H 4 ($p < 0.0001$) family gene segments less frequently (Figure 3). This pattern is due to an increase in individual gene V_H gene segment, the most prominent differences between memory IgD $-$ and plasmacytes reflected increased use of V1–2 ($p = 0.03$), V1–8 ($p = 0.003$), V2–5 ($p = 0.0003$), V3–7 ($p = 0.01$), V3–15 ($p = 0.001$), V3–30 ($p = 0.005$), and V4–40–2 ($p = 0.01$) in the former, and decreased use of V4–4 ($p = 0.0007$) and V4–34 ($p < 0.0001$) in the latter (Figure 4).

The memory IgD $-$ Ig γ repertoire used D6 ($p = 0.01$) family D_H gene segments less frequently than plasmacyte Ig γ (Figure 5). By individual D_H gene segment, the memory IgD $-$ Ig γ repertoire displayed increased use of D5–24 ($p = 0.005$) and decreased use of D2–21 ($p = 0.03$) (Figure 6). The memory IgD $-$ Ig γ repertoire used J_H 6 less frequently than plasmacytes ($p = 0.0006$) (Figure 7).

The CDR-H3 loop of the memory IgD $-$ Ig γ repertoire contained more asparagine ($p < 0.0001$) and aspartic acid ($p = 0.01$); but less tyrosine ($p = 0.04$), cysteine ($p = 0.03$), and leucine ($p = 0.01$) than plasmacyte Ig γ (Figure 8). The plasmacyte Ig γ repertoire was relatively enriched for hydrophobic amino acids, which was reflected by a higher percentage of hydrophobic CDR-H3s (hydrophobicity > 0.7) (1.54%) when compared to the memory IgD $-$ (1.12%) (Figure 9).



The Ig μ and Ig γ repertoires of analyzed cell types expressed similar distribution of D $_H$ reading frames, with reading frame 1 having greatest preference, followed by reading frame 2 and reading frame 3 (Figure 10), while the μ H chain plasmacytes used reading frame 3 less likely than memory IgD $^-$ B cells ($p = 0.03$) (Figure 10).

DISCUSSION

In both mice and humans, the composition of the antibody repertoire varies by ontogeny and by developmental stage (29, 37, 50). In order to study this process in detail, we developed a series of tools to evaluate the development of the repertoire in mice. This approach enabled us to identify constraints on V(D)J gene segment preference and CDR-H3 composition that are first established in early B cell progenitors, and then focused as the B lineage cells pass through various developmental checkpoints. The constraints are a reflection of the specific sequence from the contributing gene segments that vary in usage as a function of development (29, 30, 51–55).

Differences in the individual V-D-J gene usage, length, and amino acid composition of the adult human germline repertoires from peripheral blood and specific tissues have been previously reported (37, 50, 56–62), but comparative studies of repertoire

development in human blood have been sparse. The difficulty of study is compounded by the enhanced variability of the human repertoire when compared to mice, especially in CDR-H3. This reflects both a greater diversity of the germline sequence of the D $_H$ gene segment sequences and an increase in the extent of N addition when compared to mouse. In this work, we sought to use the same tools we had developed for the study of the mouse repertoire to perform a comparative analysis of the expressed in both the Ig μ and Ig γ repertoires in the blood of a normal, healthy human female in order to gain insight into the forces that shape the repertoire during its passage through the different stages of B cell ontogeny.

While similarities have been reported between the frequency of naïve and memory B cell repertoire usage of the V-D-J gene segments (58, 61, 62), our analysis focuses on a more detailed examination of the repertoires. Our results of low J H_1 and J H_2 usage across B cell development is consistent with previous published reports of low J H_1 and J H_2 usage in transitional, naïve, switched, and IgM memory B cell repertoires (Figure 7) (61). Altered expression of individual V H gene segments have been previously also reported in the transitional, naïve, switched, and IgM memory B cell antibody repertoires (61). As in mice, we found changes in V(D)J gene segment usage and CDR-H3 hydrophobicity in the

progression from immature to transitional to mature (**Figures 3, 5, 7, 9, and 11**). These observations support the view that the B cell receptor repertoire continues to be selected throughout early and late B cell development in the peripheral blood. Unlike mice, however, the prevalence of highly charged CDR-H3 loops increased during maturation from the immature to mature cell subsets and memory IgD[−] to plasmacyte subsets (**Figure 9**). Also unlike mice, the prevalence of highly hydrophobic CDR-H3 loops also increased in our human study subject. This may reflect a greater tolerance or preference for the use of amino acids encoded by hydrophobic D_H reading frame 2 in human B cells exposed to self and non-self antigens (35%) when compared to mice (10%), or a property specific to this particular individual, since patterns of regulation have been shown to differ in mouse strains (**Figure 10**) (34, 63).

We observed a decrease in the length of CDR-H3 during maturation (**Figures 2 and 11**). This appears to be part of a continuum of focusing CDR-H3 length in developing B cells in the bone marrow (50) and has been observed by others, as well (61). The use of long CDR-H3 loops has been previously associated with enhanced autoreactivity and polyreactivity (38, 64–66), which are presumably the features of this component of the antibody repertoire that somatic selection are designed to minimize by apoptosis or anergy.

Selection past the mature B cell stage is considered to reflect both endogenous and exogenous antigen exposure. In this regard, the most striking findings of our study were the distinctly different repertoires expressed by the memory IgD⁺ Igμ, the memory IgD[−] Igμ, and Igγ repertoires; and the plasmacyte Igμ and Igγ repertoires. We did not sort memory B cells or plasmacytes by Igμ or Igγ expression, but were able to identify unique Igμ or Igγ reads through the use of Igμ and Igγ specific primers.

The memory IgD⁺ and memory IgD[−] Igμ repertoires displayed differences in virtually all of the features of the repertoire that we evaluated, including V(D)J usage, N addition, D_H reading frame usage, CDR-H3 length, CDR-H3 loop amino acid content, and CDR-H3 hydrophobicity (**Figures 3–11**). Differences in IgD⁺ and IgD[−] Igμ repertoires in V_H1 gene family usage ($p = 0.03$) (**Figure 3**) have been reported previously (61). We observed a similar decrease in usage of V_H3–23 ($p = 0.0003$) between the memory IgD⁺ and memory IgD[−] Igμ repertoires (**Figure 4**) (61). Differences between these two memory Igμ repertoires were further enhanced by altered amino acid usage, especially an increase in arginine ($p = 0.001$) and decrease of tyrosine ($p = 0.01$) in the memory IgD[−] Igμ cell subset as compared to the memory IgD⁺ Igμ cell subset (**Figure 8**) (61). The memory IgD[−] Igμ repertoire exhibited enhanced use of charged amino acids and hydrophobic amino acids (**Figure 8**). As a result, there was a higher percentage of CDR-H3s with excess charge when compared to the memory IgD⁺ Igμ repertoire (**Figure 9**). These observations are consistent with a previous report showing that IgD⁺ memory cells had levels of negatively charged amino acids comparable to transitional and naïve B cells, while switched memory had more negatively charged residues (**Figures 8 and 9**) (61).

The vast majority of the IgD⁺ memory B cell pool also expresses IgM, whereas the IgD[−] pool expresses class-switched Ig in addition to IgM. Memory B cells expressing both IgM and IgD are considered to be the circulating equivalents of the marginal zone

B cell subset in mice; whereas memory B cells restricted to IgM production are considered to represent the more conventional B cell pool, which also is the primary source for class-switched B cells. Thus, our observations regarding the differences in repertoire between the IgD⁺ and IgD[−] memory B cell pools fit well within the view that the IgM⁺IgD⁺ and IgM⁺IgD[−] memory subsets are the products of very different immune responses. In this regard, the marginal zone-like repertoire expressed by our female study subject diverges from the marginal zone repertoire expressed in BALB/c mice in that BALB/c appears tolerant for charged CDR-H3s (35), whereas in our study subject B cells expressing charged CDR-H3s were more likely to be found in the memory IgD[−] population. Whether this difference represents a common difference between human and mouse, or reflects variation within the outbred human population is unclear and will require analysis of additional study subjects.

The plasmacyte pool represents the products of recently activated mature B cells as well as memory IgD⁺ and IgD[−] B cells that have been reactivated. This observation may explain why the plasmacyte repertoire appears intermediate between the memory IgD⁺ and IgD[−] repertoires and the mature B cell population. At present, the tools do not exist to separate plasmacytes by derivation. Moreover, the content of the memory and plasmacyte populations are likely to have been heavily influenced by several decades exposure to a variety of endogenous and exogenous antigens as well as by the anatomic niches in which the disparate subsets reside. Our study focused on bulk sequencing rather than analysis of repertoire in cells that were isolated by specific antigen reactivity, thus we cannot define the precise nature of the response to specific antigens. However, the most striking difference between the plasmacyte population and the other subsets in bulk was the decrease in the contribution of N nucleotides to the final product. Coupled with the observation that the greatest contribution of non-germline encoded nucleotides among the six subsets studied was found in the immature B cell fraction, final enrichment for germline V(D)J sequence among plasmacytes supports the view that the germline V domain repertoire has been selected by evolution for maximal advantage in responding to antigen (34–36).

As in mouse, the repertoires expressed by distinct B cell subset appear to differ in human. Sequencing of unsorted B cells from the blood is thus likely to yield an incomplete view of what is actually happening in the immune response of the individual. Our findings support the view that determination of whether diseases of immune function reflect abnormal regulation of these various B cell subsets will require considerable effort to perform deep sequencing of sorted cells from a variety of healthy individuals and patients with immune-mediated disorders (14, 38).

ACKNOWLEDGMENTS

This work was supported in part by AI090902 and AI007051.

REFERENCES

1. Hozumi N, Tonegawa S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc Natl Acad Sci U S A* (1976) 73:3628–32. doi:10.1073/pnas.73.10.3628
2. Tonegawa S. Somatic generation of antibody diversity. *Nature* (1983) 302:575–81. doi:10.1038/302575a0

3. Yancopoulos GD, Desiderio SV, Paskind M, Kearney JF, Baltimore D, Alt FW. Preferential utilization of the most JH-proximal VH gene segments in pre-B cell lines. *Nature* (1984) **311**:727–33. doi:10.1038/311727a0
4. Alt FW, Baltimore D. Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-J heavy fusions. *Proc Natl Acad Sci U S A* (1982) **79**:4118–22. doi:10.1073/pnas.79.13.4118
5. Rajewsky K. Clonal selection and learning in the antibody system. *Nature* (1996) **381**:751–8. doi:10.1038/381751a0
6. Kabat EA, Wu TT. Identical V region amino acid sequences and segments of sequences in antibodies of different specificities: relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. *J Immunol* (1991) **147**:1709–19.
7. Padlan EA. Anatomy of the antibody molecule. *Mol Immunol* (1994) **31**:169–217. doi:10.1016/0161-5890(94)90001-9
8. Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* (2000) **13**:37–45. doi:10.1016/S1074-7613(00)00006-6
9. Burrows PD, Stephan RP, Wang YH, Lassoued K, Zhang Z, Cooper MD. The transient expression of pre-B cell receptors governs B cell development. *Semin Immunol* (2002) **14**:343–9. doi:10.1016/S1044-5323(02)00067-2
10. Hardy RR, Hayakawa K. B cell development pathways. *Annu Rev Immunol* (2001) **19**:595–621. doi:10.1146/annurev.immunol.19.1.595
11. Keyna U, Beck-Engeser GB, Jongstra J, Applequist SE, Jack HM. Surrogate light chain-dependent selection of Ig heavy chain V regions. *J Immunol* (1995) **155**:5536–42.
12. Kline GH, Hartwell L, Beck-Engeser GB, Keyna U, Zaharevitz S, Klinman NR, et al. Pre-B cell receptor-mediated selection of pre-B cells synthesizing functional mu heavy chains. *J Immunol* (1998) **161**:1608–18.
13. Martin DA, Bradl H, Collins TJ, Roth E, Jack HM, Wu GE. Selection of Ig mu heavy chains by complementarity-determining region 3 length and amino acid composition. *J Immunol* (2003) **171**:4663–71.
14. Meffre E, Casellas R, Nussenzweig MC. Antibody regulation of B cell development. *Nat Immunol* (2000) **1**:379–85. doi:10.1038/80816
15. Kraus M, Alimzhanov MB, Rajewsky N, Rajewsky K. Survival of resting mature B lymphocytes depends on BCR signaling via the Ig α/β heterodimer. *Cell* (2004) **117**:787–800. doi:10.1016/j.cell.2004.05.014
16. Zikherman J, Parameswaran R, Weiss A. Endogenous antigen tunes the responsiveness of naive B cells but not T cells. *Nature* (2012) **489**:160–4. doi:10.1038/nature11311
17. Loder F, Mutschler B, Ray RJ, Paige CJ, Sideras P, Torres R, et al. B cell development in the spleen takes place in discrete steps and is determined by the quality of B cell receptor-derived signals. *J Exp Med* (1999) **190**:75–89. doi:10.1084/jem.190.1.75
18. Liu YJ, Zhang J, Lane PJ, Chan EY, MacLennan IC. Sites of specific B cell activation in primary and secondary responses to T cell-dependent and T cell-independent antigens. *Eur J Immunol* (1991) **21**:2951–62. doi:10.1002/eji.1830211209
19. Jacob J, Kassir R, Kelsoe G. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. I. The architecture and dynamics of responding cell populations. *J Exp Med* (1991) **173**:1165–75. doi:10.1084/jem.173.5.1165
20. Garside P, Ingulli E, Merica RR, Johnson JG, Noelle RJ, Jenkins MK. Visualization of specific B and T lymphocyte interactions in the lymph node. *Science* (1998) **281**:96–9. doi:10.1126/science.281.5373.96
21. Jacob K, Kelsoe G. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. II. A common clonal origin for periarteriolar lymphoid sheath-associated foci and germinal centers. *J Exp Med* (1992) **176**:679–687. doi:10.1084/jem.176.3.679
22. Klein U, Küppers R, Rajewsky K. Evidence for a large compartment of IgM-expressing memory B cells in humans. *Blood* (1997) **89**:1288–98.
23. Klein U, Rajewsky K, Küppers R. Human immunoglobulin (Ig)M+ IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells. *J Exp Med* (1998) **188**:1679–89. doi:10.1084/jem.188.9.1679
24. Tangye SG, Liu YJ, Aversa G, Phillips JH, de Vries JE. Identification of functional human splenic memory B cells by expression of CD148 and CD27. *J Exp Med* (1998) **188**:1691–703. doi:10.1084/jem.188.9.1691
25. Shi Y, Agematsu K, Ochs HD, Sugane K. Functional analysis of human memory B-cell subpopulations: IgD+ CD27+ B cells are crucial in secondary immune response by producing high affinity IgM. *Clin Immunol* (2003) **108**:128–37. doi:10.1016/S1521-6616(03)00092-5
26. Paus D, Pham RG, Chan TD, Gardam S, Basten A, Brink R. Antigen recognition strength regulates the choice between extrafollicular plasma cell and germinal center B cell differentiation. *J Exp Med* (2006) **203**:1081–91. doi:10.1084/jem.20060087
27. Berek C, Berger A, Apel M. Maturation of the immune response in germinal centers. *Cell* (1991) **67**:1121–9. doi:10.1016/0092-8674(91)90289-B
28. Dal Porto JM, Haberman AM, Shlomchik MJ, Kelsoe G. Antigen drives very low affinity B cells to become plasmacytes and enter germinal centers. *J Immunol* (1998) **161**:5373–81.
29. Schroeder HW Jr, Ippolito GC, Shiokawa S. Regulation of the antibody repertoire through control of HCDR3 diversity. *Vaccine* (1998) **16**:1383–90. doi:10.1016/S0264-410X(98)00096-6
30. Ippolito GC, Schelonka RL, Zemlin M, Ivanov II, Kobayashi R, Zemlin C, et al. Forced usage of positively charged amino acids in immunoglobulin CDR-H3 impairs B cell development and antibody production. *J Exp Med* (2006) **203**:1567–78. doi:10.1084/jem.20052217
31. Ivanov II, Schelonka RL, Zhuang Y, Gartland GL, Zemlin M, Schroeder HW Jr. Development of the expressed immunoglobulin CDR-H3 repertoire is marked by focusing of constraints in length, amino acid utilization, and charge that are first established in early B cell progenitors. *J Immunol* (2005) **174**:7773–80.
32. Schelonka RL, Ivanov II, Jung DH, Ippolito GC, Nitschke K, Zhuang Y, et al. A single DH gene segment creates its own unique CDR-H3 repertoire and is sufficient for B cell development and immune function. *J Immunol* (2005) **175**(10):6624–32.
33. Schelonka RL, Zemlin M, Kobayashi R, Szalai A, Ippolito GC, Zhuang Y, et al. Preferential use of DH reading frame 2 alters B cell development and antigen-specific antibody production. *J Immunol* (2008) **181**:8409–15.
34. Khass M, Buckley K, Kapoor P, Schelonka RL, Watkins LS, Zhuang Y, et al. Recirculating bone marrow B cells in C57BL/6 mice are more tolerant of highly hydrophobic and highly charged CDR-H3s than those in BALB/c mice. *Eur J Immunol* (2013) **43**(3):629–40. doi:10.1002/eji.201242936
35. Raaphorst FM, Raman CS, Tami J, Fischbach M, Sanz I. Human Ig heavy chain CDR3 regions in adult bone marrow pre-B cells display an adult phenotype of diversity: evidence for structural selection of DH amino acid sequences. *Int Immunol* (1997) **9**:1503–15. doi:10.1093/intimm/9.10.1503
36. Schroeder HW Jr, Zemlin M, Khass M, Nguyen HH, Schelonka RL. Genetic control of DH reading frame and its effect on B-cell development and antigen-specific antibody production. *Crit Rev Immunol* (2010) **30**:327–44. doi:10.1615/CritRevImmunol.v30.i4.20
37. Ivanov II, Link JM, Ippolito GC, Schroeder HW Jr. Constraints on hydrophobicity and sequence composition of HCDR3 are conserved across evolution. In: Zanetti M, Capra JD, editors. *The Antibodies*. London: Taylor and Francis Group (2002). p. 43–67.
38. Wardemann H, Yurasov S, Schaefer A, Young JW, Meffre E, Nussenzweig MC. Predominant autoantibody production by early human B cell precursors. *Science* (2003) **301**:1374–7. doi:10.1126/science.1086907
39. Lim PS, Shannon MF, Hardy K. Epigenetic control of inducible gene expression in the immune system. *Epigenomics* (2010) **2**:775–95. doi:10.2217/epi.10.55
40. Ippolito GC, Hoi KH, Reddy ST, Carroll SM, Ge X, Rogosch T, et al. Antibody repertoires in humanized NOD-SCID-IL2Rγ(null) mice and human B cells reveals human-like diversification and tolerance checkpoints in the mouse. *PLoS One* (2012) **7**:e35497. doi:10.1371/journal.pone.0035497
41. Brochet X, Lefranc MP, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* (2008) **36**:W503–8. doi:10.1093/nar/gkn316
42. Rogosch T, Kerzel S, Hoi KH, Zhang Z, Maier RF, Ippolito GC, et al. Immunoglobulin analysis tool: a novel tool for the analysis of human and mouse heavy and light chain transcripts. *Front Immunol* (2012) **3**:1–14. doi:10.3389/fimmu.2012.00176
43. Foerster C, Voelken N, Rakhamov M, Keller B, Gutenberger S, Goldacker S, et al. B cell receptor mediated calcium signaling is impaired in B lymphocytes of Type Ia patients with common variable immunodeficiency. *J Immunol*. (2010) **184**:7305–13. doi:10.4049/jimmunol.1000434

44. Manjarrez-Orduno N, Quach TD, Sanz I. B cells and immunological tolerance. *J Invest Dermatol* (2009) **129**:278–88. doi:10.1038/jid.2008.240
45. Palanichamy A, Barnard J, Zheng B, Owen T, Quach T, Wei C, et al. Novel human transitional B cell populations revealed by B cell depletion therapy. *J Immunol* (2009) **182**:7982–3. doi:10.4049/jimmunol.0801859
46. Sanz I, Wei C, Lee FEH, Anolik J. Phenotypic and functional heterogeneity of human memory B cells. *Semin Immunol* (2008) **20**:67–82. doi:10.1016/j.smim.2007.12.006
47. Warnatz K, Denz A, Drager R, Braun M, Groth C, Wolff-Vorbeck G, et al. Severe deficiency of switched memory B cells (CD27+IgM-IgD-) in subgroups of patients with common variable immunodeficiency: a new approach to classify a heterogeneous disease. *Blood* (2002) **99**(5):1544–51. doi:10.1182/blood.V99.5.1544
48. Kyte J, Doolittle RF. A simple method for displaying the hydrophobic character of a protein. *J Mol Biol* (1982) **157**:105–32. doi:10.1016/0022-2836(82)90515-0
49. Eisenberg D. Three-dimensional structure of membrane and surface proteins. *Annu Rev Biochem* (1984) **53**:595–623. doi:10.1146/annurev.bi.53.070184.003115
50. Schroeder HW Jr. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev Comp Immunol* (2005) **30**:119–35. doi:10.1016/j.dci.2005.06.006
51. Asma GE, van den Berg RL, Vossen JM. Characterization of early lymphoid precursor cells in the human fetus using monoclonal antibodies and anti-terminal deoxynucleotidyl transferase. *Clin Exp Immunol* (1986) **64**:356–63.
52. Logtenberg T, Schutte MEM, Ebeling SB, Gmelig-Meyling FHJ, Van Es JH. Molecular approaches to the study of human B-cell and (auto)antibody repertoire generation and selection. *Immunol Rev* (1992) **128**:23–47. doi:10.1111/j.1600-065X.1992.tb00831.x
53. Schelonka RL, Ivanov II, Jung D, Ippolito GC, Nitschke L, Zhuang Y, et al. A single DH gene segment is sufficient for B cell development and immune function. *J Immunol* (2005) **175**:6624–32.
54. Terrell TG, Holmberg CA, Osburn BI. Immunologic surface markers on non-human primate lymphocytes. *Am J Vet Res* (1977) **38**:503–7.
55. Zemlin M, Schelonka RL, Ippolito GC, Zemlin C, Zhuang Y, Gartland GL, et al. Regulation of repertoire development through genetic control of DH reading frame preferences. *J Immunol* (2008) **181**:8416–24.
56. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-resolution description of antibody heavy chain repertoires in humans. *PLoS One* (2011) **6**:e22365. doi:10.1371/journal.pone.0022365
57. Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* (2010) **184**:6986–92. doi:10.4049/jimmunol.1000445
58. Briney BS, Willis JR, McKinney BA, Crowe JE Jr. High-throughput antibody sequencing reveals genetic evidence of global regulation of the naïve and memory repertoires that extends across individuals. *Genes Immun* (2012) **13**:469–73. doi:10.1038/gene.2012.20
59. Kraj P, Friedman DF, Stevenson F, Silberstein LE. Evidence for the overexpression of the VH4-34 (VH4.21) Ig gene segment in the normal adult human peripheral blood B cell repertoire. *J Immunol* (1995) **154**:6406–20.
60. Stewart AK, Huang C, Stollar BD, Schwartz RS. High-frequency representation of a single VH gene in the expressed human B cell repertoire. *J Exp Med* (1993) **177**:409–18. doi:10.1084/jem.177.2.409
61. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B cell populations. *Blood* (2010) **116**:1070–8. doi:10.1182/blood-2010-03-275859
62. Larimore K, McCormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH repertoires revealed by deep sequencing. *J Immunol* (2012) **189**:3221–30. doi:10.4049/jimmunol.1201303
63. Zemlin M, Ippolito GC, Zemlin C, Link J, Monestier M, Schroeder HW Jr. Adult lupus-prone MRL/Mpj2+ mice express a primary antibody repertoire that differs in CDR-H3 length distribution and hydrophobicity from that expressed in the C3H parental strain. *Mol Immunol* (2005) **42**(7):789–98. doi:10.1016/j.molimm.2004.07.049
64. Aguilera I, Melero J, Nunez-Roldan A, Sanchez B. Molecular structure of eight human autoreactive monoclonal antibodies. *Immunology* (2001) **102**:273–80. doi:10.1046/j.1365-2567.2001.01159.x
65. Ichiyoshi Y, Casali P. Analysis of the structural correlates for antibody polyreactivity by multiple reassortments of chimeric human immunoglobulin heavy and light chain V segments. *J Exp Med* (1994) **180**:885–95. doi:10.1084/jem.180.3.885
66. Klonowski KD, Primiano LL, Monestier M. Atypical VH-D-JH rearrangements in newborn autoimmune MRL mice. *J Immunol* (1999) **162**:1566–72.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 November 2013; accepted: 23 February 2014; published online: 19 March 2014.

*Citation: Mrocze ES, Ippolito GC, Rogosch T, Hoi KH, Hwangpo TA, Brand MG, Zhuang Y, Liu CR, Schneider DA, Zemlin M, Brown EE, Georgiou G and Schroeder HW Jr. (2014) Differences in the composition of the human antibody repertoire by B cell subsets in the blood. *Front. Immunol.* **5**:96. doi: 10.3389/fimmu.2014.00096*

This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.

Copyright © 2014 Mrocze, Ippolito, Rogosch, Hoi, Hwangpo, Brand, Zhuang, Liu, Schneider, Zemlin, Brown, Georgiou and Schroeder. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Secondary mechanisms of diversification in the human antibody repertoire

Bryan S. Briney¹ and James E. Crowe Jr.^{1,2,3*}

¹ Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, USA

² Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA

³ The Vanderbilt Vaccine Center, Vanderbilt University Medical Center, Nashville, TN, USA

Edited by:

Harry W. Schroeder, University of Alabama at Birmingham, USA

Reviewed by:

Robert A. Eisenberg, University of Pennsylvania, USA

To-Ha Thai, Beth Deaconess Israel Medical Center, USA

***Correspondence:**

James E. Crowe Jr., Vanderbilt Vaccine Center, Vanderbilt University Medical Center, 11475 MRB IV, 2213 Garland Avenue, Nashville, TN 37232-0417, USA.
e-mail: james.crowe@vanderbilt.edu

V(D)J recombination and somatic hypermutation (SHM) are the primary mechanisms for diversification of the human antibody repertoire. These mechanisms allow for rapid humoral immune responses to a wide range of pathogenic challenges. V(D)J recombination efficiently generate a virtually limitless diversity through random recombination of variable (V), diversity (D), and joining (J) genes with diverse non-templated junctions between the selected gene segments. Following antigen stimulation, affinity maturation by SHM produces antibodies with refined specificity mediated by mutations typically focused in complementarity determining regions (CDRs), which form the bulk of the antigen recognition site. While V(D)J recombination and SHM are responsible for much of the diversity of the antibody repertoire, there are several secondary mechanisms that, while less frequent, make substantial contributions to antibody diversity including V(DD)J recombination (or D-D fusion), SHM-associated insertions and deletions, and affinity maturation and antigen contact by non-CDR regions of the antibody. In addition to enhanced diversity, these mechanisms allow the production of antibodies that are critical to response to a variety of viral and bacterial pathogens but that would be difficult to generate using only the primary mechanisms of diversification.

Keywords: VDJ rearrangement, VH replacement, B Cell Biology, human, insertion/deletion

INTRODUCTION

A diverse antibody repertoire is a principal component of humoral immunity and is critical to the development of functional adaptive immune responses. Generation of this repertoire diversity is accomplished primarily through two mechanisms: recombination and somatic hypermutation (SHM). These two mechanisms produce massive diversity within antibody complementarity determining regions (CDRs), which form the primary antigen contact site. The availability of multiple variable genes for selection at the time of recombination facilitates large combinatorial diversity, which is further expanded by a diversity of possible heavy and light chain combinations. In this review, we discuss in detail three additional mechanisms which, while less common than recombination and SHM, contribute substantially to the generation of diversity within the antibody repertoire: (1) non-standard recombinations that violate the 12/23 rule of recombination, (2) SHM-associated genetic insertions and deletions, and (3) affinity maturation and direct antigen contact by non-CDR antibody regions.

V(D)J RECOMBINATION: FOLLOWING THE 12/23 RULE

Since the discovery that recombination activating gene (RAG)-mediated recombination of variable (V), diversity (D) and joining (J) genes generates virtually unlimited sequence diversity in the antibody repertoire (Brack et al., 1978; Alt and Baltimore, 1982; Tonegawa, 1983; Schatz et al., 1989; Oettinger et al., 1990), much progress has been made in determining the genetic and mechanistic elements that participate in the antibody recombination process. It is generally understood that recombination signal

sequences (RSS), which are composed of conserved AT-rich heptamer and nonamer sequences separated by spacers of either 12 or 23 nucleotides, are recognized and bound by RAG1 and RAG2 proteins at the initiation of the recombination process (Hesse et al., 1989; Alt et al., 1992). RAG binding is highly dependent on the heptamer and nonamer sequences, and alterations to either sequence results in decreased RAG binding (Cuomo et al., 1996; Difilippantonio et al., 1996; Nadel et al., 1998). The length of the spacer sequence is critical to recombination, and there is evidence of sequence conservation within the spacer region (Ramsden et al., 1994; Lee et al., 2003; Montalbano et al., 2003).

Recombination typically occurs only between RSS elements of different spacer lengths, in a model commonly referred to as the 12/23 rule of recombination (Ramsden et al., 1996; Steen et al., 1996; van Gent et al., 1996; Schatz, 2004). After binding to one 12-bp RSS and one 23-bp RSS, the RAG complex induces single-strand DNA nicks between the coding sequence and the heptamer of each RSS, resulting in hairpin formation on each of the coding ends and a blunt double-stranded break on each signal end (Roth et al., 1992; Schlissel et al., 1993; McBlane et al., 1995; Sadofsky, 2001). The hairpins are opened, nucleotides may be added to or removed from the coding ends, and the double-strand DNA breaks at the coding ends are joined into a single coding strand (Lewis, 1994; Mahajan et al., 1999; Shockett and Schatz, 1999; Walker et al., 2001; Mansilla-Soto and Cortes, 2003; Roth, 2003).

In antibody heavy chain genes, D gene segments are flanked by 12-bp RSSs on either side, while V_H and J_H gene segments are flanked by 23-bp RSSs (Early et al., 1980;

Kurosawa and Tonegawa, 1982). Recombination thus proceeds in a step-wise fashion, with D–J_H recombination preceding V_H–D recombination, resulting in a complete heavy chain variable region (Alt et al., 1987; Schatz et al., 1992). A single recombination event joins the light chain V and J gene, and pairing of recombined heavy chain and recombined light chains results in massive diversity within the unmutated antibody repertoire.

NON-12/23 RECOMBINATION: V(DD)J AND DIRECT V_H–J_H RECOMBINATION

Direct V_H–J_H joining and V(DD)J recombination (also referred to as D–D fusion) are in direct violation of the 12/23 rule, but such recombination events have been demonstrated in both *in vitro* and *in vivo* systems (Sanz, 1991; Kiyoi et al., 1992; Raaphorst et al., 1997; Koralov et al., 2005, 2006; Watson et al., 2006). Even in model systems designed to induce such recombination events, however, non-12/23 recombinations are much less efficient than recombinations that adhere to the 12/23 rule (Akira et al., 1987; Hesse et al., 1989; Akamatsu et al., 1994).

V(DD)J recombinants are the result of an aberrant recombination process by which two or more D genes are joined into a single recombinant. The joining of two D genes, which are flanked on both sides by 12-bp RSSs, can only be accomplished in clear violation of the 12/23 rule, but recombined antibody genes in this configuration have now been isolated by numerous investigators. While V(DD)J recombination typically results in an unusually long heavy chain CDR 3 (HCDR3) region, the use of two D segments is not the primary mechanism by which long HCDR3 loops are generated (Briney et al., 2012a). Long HCDR3s typically are generated by the use of longer D and J segments and long non-templated junctional regions. The precise order of events during the V(DD)J recombination process is unclear: it is not known whether V(DD)J recombinants are produced through an additional D–D recombination following the initial D–J_H recombination, or whether D–D fusion occurs before, even long before, the D–J_H recombination. V(DD)J recombinations have been estimated by some to occur in as many as 5–11% of all recombinations (Sanz, 1991; Kiyoi et al., 1992; Raaphorst et al., 1997), but the true frequency of V(DD)J recombinations is difficult to determine. Identification of V(DD)J recombinants relies on the accurate detection of two diversity genes within a single recombinant, but N-addition mimicry of diversity gene segments, which is genetically indistinguishable from true V(DD)J recombination, likely inflates many published estimates of V(DD)J recombination (Watson et al., 2006). Recent work, which leveraged high-throughput sequencing and a stringent filtering process, placed a lower bound of the frequency of V(DD)J recombinants in the human peripheral blood repertoire at approximately 1 in 800 B cells (Briney et al., 2012b).

The occurrence of direct V_H–J_H recombination, like V(DD)J recombination, requires clear violation of the 12/23 rule, since both V_H and J_H segments are flanked by 23-bp RSSs. Little is known about the frequency of direct V_H–J_H recombination in the human repertoire. Several studies of the human CDR3 repertoire that have identified D–D fusions have failed to identify V_H–J_H recombinants, indicating that if they occur, V_H–J_H recombinations are likely very rare (Sanz, 1991; Kiyoi et al.,

1992; Raaphorst et al., 1997; Watson et al., 2006). This finding is somewhat surprising, since *in vitro* recombination between two 23-bp RSSs occurred much more frequently than recombination between two 12-bp RSSs (Jones and Gellert, 2002). In contrast to D–D fusions, for which there are several studies on the frequency of V(DD)J recombinants in the human peripheral blood repertoire, much of the published work describing *in vivo* V_H–J_H recombination relies on transgenic mouse models lacking D gene loci (Koralov et al., 2005, 2006). Since these model systems produce only aberrant recombinants, it is difficult to interpret the resulting data in terms of the likely occurrence and frequency of such recombinants in the naturally occurring circulating B cell repertoire. As with V(DD)J recombination, determination of the true frequency of direct V_H–J_H recombination will likely prove difficult, as extensive chewback of D genes during normal V(D)J recombination may appear genetically indistinguishable from true V_H–J_H recombination and inflate any estimates of the frequency of V_H–J_H recombination.

NON-12/23 RECOMBINATION: V_H REPLACEMENT AND RECEPTOR REVISION

V_H replacement is a process by which a secondary V_H–V(D)J recombination can occur, resulting in replacement of the variable gene while preserving the original D–J_H recombination. V_H replacement, which is thought to be a form of heavy chain receptor editing, differs from light chain receptor editing, although both typically occur early in B cell development (Prak and Weigert, 1995; Nemazee and Weigert, 2000). Light chain receptor editing results in an entirely new V_L–J_L recombination through the recombination of a V_L gene segment upstream of the original recombination with a J_L gene segment downstream of the original recombination (Papavasiliou et al., 1997; Retter and Nemazee, 1998). Thus, light chain receptor editing proceeds without violating the 12/23 rule. In contrast, V_H replacement involves V_H–V(D)J recombination, which results in retention of the original D–J_H junction and replacement only of the V_H gene segment (Kleinfield and Weigert, 1989; Nemazee, 2006). V_H replacement utilizes a cryptic RSS (cRSS) found near the 3' end of most human variable genes (Radic and Zouali, 1996), and this cRSS is used to recombine with the normal RSS at the 3' end of the invading variable gene. The cRSS contains a heptamer sequence, but lacks an identifiable nonamer or spacer sequence, and recombination with the cRSS is inefficient, much like other forms of non-12/23 recombination (Koralov et al., 2006; Lutz et al., 2006).

V_H replacement also can be distinguished from receptor revision, which is putatively antigen-driven and has not been shown to use the conserved cRSS elements near the 3' end of the V gene. Instead, receptor revisions are suggested to occur peripherally in mature B cells using alternate RSS-like elements that sometimes contain only the CAC motif found at the 5' end of most RSS heptamers or the inverse GTG motif found at the 3' end; the few examples of this phenomenon typically occurred near the middle of heavy chain framework region (FR) 3 (Itoh et al., 2000; Wilson et al., 2000; Lenze et al., 2003). Use of these alternate RSS-like elements results in formation of a hybrid V gene, retaining a substantial portion of the initially recombined V gene,

as opposed to the nearly complete removal of the initially recombinant V gene observed in V_H replacement. Because the observed receptor revision events occurred in stretches of sequence similarity between V genes, it has been proposed that these revisions may instead be polymerase chain reaction (PCR) artifacts caused by incomplete recombinant amplification followed by priming of a different V(D)J recombinant with the partially amplified fragment, resulting in a hybrid sequence (Darlow and Stott, 2005). In approximately half of all identified receptor revisions in these studies, the invading V gene is located downstream of the variable gene used in the initial V(D)J recombination, which would not be possible using the proposed receptor revision mechanism. Inter-chromosomal recombination has been proposed as the mechanism for these out-of-order receptor revisions (Wilson et al., 2000). More recent work has shown that receptor reverions are not observed when amplifying from single B cells (Goossens et al., 2001), providing further evidence that the previously observed receptor revisions may be an artifact of PCR amplification of multiple antibody sequences from bulk B cells.

It is thought that V_H replacement, like other forms of receptor editing, occurs primarily in the immature B cell population to rescue non-functional or autoreactive recombinants (Zhang et al., 2004; Lutz et al., 2006), but some studies suggest that V_H replacement may be possible in mature B cells (Hikida et al., 1996; Han et al., 1997; Papavasiliou et al., 1997; Hertz et al., 1998; Nussen-zweig, 1998). Somewhat paradoxically, V_H replacement, which is purported to be a primary mechanism for resolving self-reactive recombinations, can itself result in antibodies with autoreactive characteristics (Klonowski and Monestier, 2000; Zhang et al., 2003). V_H replacement was observed first in transformed murine pre-B cells (Kleinfield et al., 1986; Reth et al., 1986), with subsequent studies identifying V_H replacement *in vivo* (Taki et al., 1993; Chen et al., 1995). In the most informative work done on V_H replacement in the human repertoire, a genetic fingerprint of V_H replacement was identified in the human peripheral blood repertoire (Zhang et al., 2003). Identification of V_H replacement events in the peripheral repertoire relies on detection of short pentameric sequences that are located between the cRSS and the 3' end of V genes. These pentamers remain even after V_H replacement, providing an identifiable remnant of the replaced V gene. Short pentameric sequences are easily mimicked through random N-addition, making reliable detection of V_H replacement difficult. Therefore, estimates of V_H recombination frequency in the peripheral blood repertoire have varied widely, from 5 to 22% of the total repertoire (Zhang et al., 2003; Koralov et al., 2006; Watson et al., 2006).

SOMATIC HYPERMUTATION

In humans and in mice, diversification of the secondary antibody repertoire, which arises in response to antigenic stimulus, is accomplished primarily through SHM (Brenner and Milstein, 1966; Kelsoe, 1994). Naïve, antigen-inexperienced B cells undergo the SHM process upon recognition of an infectious agent. It is through the SHM process, which occurs primarily in secondary lymphoid tissue, that hosts mutate the variable region of their antibody genes (MacLennan et al., 1992; Li et al., 2004). Many of these mutations have no effect on antigen recognition and many

have deleterious effects on either antigen recognition or proper folding of the antibody protein. Some mutations, however, produce antibodies with improved affinity for the target pathogenic epitope (Casali et al., 2006). Thus, the SHM process provides a basis for the positive selection of high-affinity antibodies that are characteristic of a mature immune response (MacLennan, 1994).

Many components of the SHM machinery are known, but the complete process and the mechanisms by which it is targeted specifically to the immunoglobulin loci are still poorly understood. SHM introduces point mutations at a frequency of approximately 10^{-3} mutations per base pair, which is about 10^6 -fold higher than the rate of spontaneous mutation in other genes (Rajewsky et al., 1987). Mutations begin approximately 150-bp downstream of the transcription start site and the mutation frequency decreases exponentially with increasing distance from the transcription start site (Rada and Milstein, 2001). Activation-induced cytidine deaminase (AID) is required for SHM and initiates the SHM process by the deamination of C nucleotides (Muramatsu et al., 1999, 2000). Deamination results in a U–G mismatch, and several possible processes result in the error-prone repair of the mismatch. Although the precise mechanism(s) responsible for error-prone repair during SHM are not known, several DNA repair mechanisms have been shown to be critical to the SHM process, including base excision repair and mismatch repair (Phung et al., 1998; Rada et al., 1998; Wiesendanger et al., 2000; Di Noia and Neuberger, 2002; Zheng et al., 2005).

SOMATIC HYPERMUTATION-ASSOCIATED INSERTIONS AND DELETIONS

Although the SHM process typically results in single nucleotide substitutions, deletion of germline nucleic acids or insertion of non-germline nucleic acids does occur in association with SHM (Goossens et al., 1998; Wilson et al., 1998a; Bemark and Neuberger, 2003). These insertions and deletions (indels) are rare, with SHM-associated (SHA) indels estimated to be present in 1.3–6.5% of circulating B cells (Goossens et al., 1998; Wilson et al., 1998a; Bemark and Neuberger, 2003). Short SHA indels are much more common than long SHA indels, with most insertions and deletions being 1–2 codons in length (Goossens et al., 1998; Wilson et al., 1998a; Bemark and Neuberger, 2003). Although infrequent, SHA insertion and deletion events add substantially to the diversity of the human antibody repertoire (Wilson et al., 1998b; de Wildt et al., 1999; Reason and Zhou, 2006).

Somatic hypermutation-associated insertions and deletions also have been shown to play a critical role in the antibody response against viral and bacterial pathogens, including HIV, influenza, and *Streptococcus pneumoniae* (Zhou et al., 2004; Walker et al., 2009, 2011; Wu et al., 2010a; Krause et al., 2011; Pejchal et al., 2011). Of particular interest, structural analysis of an SHA insertion in the anti-influenza antibody 2D1 identified a substantial structural alteration induced by the insertion (Krause et al., 2011). This insertion, although located in a FRs, caused a large conformational change in a CDR and allowed antibody–antigen interactions that were sterically hindered without the insertion-induced conformational change. In addition to 2D1, the extremely broad and potently neutralizing HIV antibody VRC01 contained a six nucleotide deletion in the CDR1 of the light

chain (CDR-L1; Wu et al., 2010a). This SHA deletion shortened the CDR-L1 loop, thereby removing potential clashes with loop D of the HIV envelope protein and allowing direct interaction between the HIV antigen and the CDR-L2 loop of VRC01 (Zhou et al., 2010).

ANTIBODY COMPLEMENTARITY DETERMINING REGIONS

Antibody CDRs (also referred to as hypervariable regions) are the primary region of antigen recognition, contain extensive sequence diversity even among germline genes, and are targeted preferentially for affinity maturation, making them the most variable regions of the antibody gene (Capra and Kehoe, 1975; Kabat et al., 1992). There are several structural and genetic reasons for the preferential targeting of CDRs by SHM. Genetically, SHM is known to preferentially target the WRCY hotspot motif (or its reverse complement, RGYW; Dörner et al., 1998), and the frequency of these hotspots is increased in CDRs (Wagner et al., 1995; Shapiro and Wysocki, 2002; Pham et al., 2003). Further, codon usage is biased in CDRs toward codons that are easily mutable, enhancing the likelihood that a nucleotide substitution induced by SHM results in an amino acid change (Motoyama et al., 1991; Wagner et al., 1995; Kepler, 1997). Structurally, the CDRs are largely loop-based, which make them sufficiently flexible to incorporate the substitutions and short indels introduced by SHM without compromising structural integrity. FRs, by contrast, are highly structured and less able to accommodate somatic mutations (Celada and Seiden, 1996).

AFFINITY MATURATION AND ANTIGEN CONTACT BY ANTIBODY FRAMEWORK REGIONS

While much affinity maturation is focused on the CDRs, there are other regions that are important to antigen recognition. T cell receptors (TCRs) contain a fourth hypervariable region (HV4, sometimes referred to as CDR4), which is highly variable, surface-exposed, and involved in superantigen and accessory molecule recognition (Choi et al., 1990; Garcia et al., 1996; Li et al., 1998). We have recently used high-throughput sequencing approaches to determine the sequence of thousands of antibody genes containing SHM-associated insertions and deletions (SHA indels), which revealed significant differences between the location of SHA indels and somatic mutations (Briney et al., 2012c). Further, we identified a cluster of insertions and deletions in the antibody FR3 region that corresponds to the HV4 in TCRs.

Emerging evidence suggests that an HV4-like region may exist in antibodies as well as TCRs. Recent crystallographic work on the anti-influenza antibody CR6261 has shown that the HV4-like region of FR3 was somatically mutated (Throsby et al., 2008) and directly contributed to antigen binding (Ekiert et al., 2009). The anti-influenza antibody 2D1 contains a three-codon insertion in a HV4-like region of FR3 which, while not directly involved in antigen recognition, causes a critical conformational shift in nearby CDRs that is required for antigen recognition (Krause et al., 2011). A unique example of HV4-like contribution to antigen recognition is the anti-HIV antibody 21c (Diskin et al., 2010). 21c binds to the HIV co-receptor binding pocket, which is only exposed following binding of CD4, the primary host receptor. Interestingly, while the majority of the binding surface of 21c is in contact

with the HIV envelope protein, the HV4-like region of 21c binds to CD4, forming a cross-protein epitope. In addition to 21c, the broadly neutralizing anti-HIV antibody VRC03 contains a surprisingly long seven-codon insertion in the HV4-like region of FR3 (Wu et al., 2010a). Finally, the HV4-like FR3 region of antibody heavy chains of the V_H3 family has been shown to interact with Staphylococcal protein A, a known superantigen (Potter et al., 1996), mimicking the superantigen-binding activity of the HV4 region in TCRs. While the HV4-like regions that have been identified to date are not somatically mutated to the same extent as antibody CDRs, the ability of this HV4-like region to tolerate a substantial number of somatic mutations and genetic insertions suggests the existence of a somewhat flexible region that has an under-appreciated ability to accommodate affinity maturation modifications.

CONCLUSION

V(DD)J recombination, SHA indels, and antigen contact by non-CDR antibody regions, while secondary to V(D)J recombination and SHM as mechanisms of antibody diversification, contribute substantially to antibody diversity. Each of these secondary affinity maturation mechanisms allows for the generation of unique genetic or structural elements that have been shown to be important to the humoral response against a variety of viral and bacterial pathogens including HIV, influenza virus, staphylococci and pneumococci. These secondary affinity maturation events are much less common than SHM and, as a consequence, are more difficult to study effectively. The advent of next-generation sequencing technology has made it possible to obtain thousands or millions, and soon to be billions, of antibody sequences (Boyd et al., 2009, 2010; Wu et al., 2010b; Prabakaran et al., 2011; Briney et al., 2012d). It is likely that over the coming years, this digital flood of antibody sequence data will allow a much more complete understanding of these secondary affinity maturation events. For example, current technologies for isolating antigen-specific antibodies from human blood or bone marrow cells are relatively inefficient and result in stochastic discovery of unique antibodies. High-throughput sequence analysis techniques now allow comprehensive definition of all expressed antibody sequences in samples, even to the scale of analyzing all antibody sequences in leukopacks containing most of the circulating B cells in an individual at a time point. Novel methods under current development for determining phylogenetic relationships among expressed antibody sequences may allow us to define the path of somatic mutation from unmutated ancestor sequences to the final affinity-matured antigen-specific sequence. Likely, these studies will reveal that B cell clones that develop following antigen stimulation do not follow linear paths of development, but rather diverge into complex families with multiply branched phylogenies. Such studies should greatly broaden our understanding of the molecular and genetic events occurring in the B cell repertoire following antigen stimulation.

ACKNOWLEDGMENT

This work was supported by U01 AI-078407 (NIAID, NIH), NIAID Contract HHSN272200900047C and DoD grant HDTRA1-10-1-0067.

REFERENCES

- Akamatsu, Y., Tsurushita, N., Nagawa, F., Matsuoka, M., Okazaki, K., Imai, M., et al. (1994). Essential residues in V(D)J recombination signals. *J. Immunol.* 153, 4520–4529.
- Akira, S., Okazaki, K., and Sakano, H. (1987). Two pairs of recombination signals are sufficient to cause immunoglobulin V-(D)-J joining. *Science* 238, 1134–1138.
- Alt, F., Blackwell, T., and Yancopoulos, G. (1987). Development of the primary antibody repertoire. *Science* 238, 1079–1087.
- Alt, F. W., and Baltimore, D. (1982). Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-J_H fusions. *Proc. Natl. Acad. Sci. U.S.A.* 79, 4118–4122.
- Alt, F. W., Oltz, E. M., Young, F., Gorman, J., Taccioli, G., and Chen, J. (1992). VDJ recombination. *Immunol. Today* 13, 306–314.
- Bemark, M., and Neuberger, M. S. (2003). By-products of immunoglobulin somatic hypermutation. *Genes Chromosomes Cancer* 38, 32–39.
- Boyd, S. D., Gaëta, B. A., Jackson, K. J., Fire, A. Z., Marshall, E. L., Merker, J. D., et al. (2010). Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* 184, 6986–6992.
- Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., et al. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.* 1, 12ra23.
- Brack, C., Hirama, M., Lenhard-Schuller, R., and Tonegawa, S. (1978). A complete immunoglobulin gene is created by somatic recombination. *Cell* 15, 1–14.
- Brenner, S., and Milstein, C. (1966). Origin of antibody variation. *Nature* 211, 242–243.
- Briney, B. S., Willis, J. R., and Crowe, J. E. (2012a). Human peripheral blood antibodies with long HCDR3s are established primarily at original recombination using a limited subset of germline genes. *PLoS ONE* 7:e36750. doi: 10.1371/journal.pone.0036750
- Briney, B. S., Willis, J. R., Hicar, M. D., Thomas, J. W., and Crowe, J. E. (2012b). Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology* 137, 56–64.
- Briney, B. S., Willis, J. R., and Crowe, J. E. Jr. (2012c). Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. *Genes Immun.* 13, 523–529.
- Briney, B. S., Willis, J. R., McKinney, B. A., and Crowe, J. E. Jr. (2012d). High-throughput antibody sequencing reveals genetic evidence of global regulation of the naïve and memory repertoires that extends across individuals. *Genes Immun.* 13, 469–473.
- Capra, J. D., and Kehoe, J. M. (1975). Hypervariable regions, idiotype, and the antibody-combining site. *Adv. Immunol.* 20, 1–40.
- Casali, P., Pal, Z., Xu, Z., and Zan, H. (2006). DNA repair in antibody somatic hypermutation. *Trends Immunol.* 27, 313–321.
- Celada, F., and Seiden, P. E. (1996). Affinity maturation and hypermutation in a simulation of the humoral immune response. *Eur. J. Immunol.* 26, 1350–1358.
- Chen, C., Nagy, Z., Prak, E., and Weigert, M. (1995). Immunoglobulin heavy chain gene replacement: a mechanism of receptor editing. *Immunity* 3, 747–755.
- Choi, Y. W., Herman, A., Digiusto, D., Wade, T., Marrack, P., and Kappler, J. (1990). Residues of the variable region of the T-cell-receptor beta-chain that interact with *S. aureus* toxin superantigens. *Nature* 346, 471–473.
- Cuomo, C. A., Mundy, C. L., and Oettinger, M. A. (1996). DNA sequence and structure requirements for cleavage of V(D)J recombination signal sequences. *Mol. Cell. Biol.* 16, 5683–5690.
- Darlow, J. M., and Stott, D. I. (2005). V(H) replacement in rearranged immunoglobulin genes. *Immunology* 114, 155–165.
- de Wildt, R. M., van Venrooij, W. J., Winter, G., Hoet, R. M., and Tomlinson, I. M. (1999). Somatic insertions and deletions shape the human antibody repertoire. *J. Mol. Biol.* 294, 701–710.
- Difilippantonio, M. J., McMahan, C. J., Eastman, Q. M., Spanopoulou, E., and Schatz, D. G. (1996). RAG1 mediates signal sequence recognition and recruitment of RAG2 in V(D)J recombination. *Cell* 87, 253–262.
- Di Noia, J., and Neuberger, M. S. (2002). Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. *Nature* 419, 43–48.
- Diskin, R., Marcovecchio, P. M., and Bjorkman, P. J. (2010). Structure of a clade C HIV-1 gp120 bound to CD4 and CD4-induced antibody reveals anti-CD4 polyreactivity. *Nat. Struct. Mol. Biol.* 17, 608–613.
- Dörner, T., Foster, S., Farmer, N., and Lipsky, P. (1998). Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. *Eur. J. Immunol.* 28, 3384–3396.
- Early, P., Huang, H., Davis, M., Calame, K., and Hood, L. (1980). An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. *Cell* 19, 981–992.
- Ekiert, D. C., Bhabha, G., Elsliger, M.-A., Friesen, R. H. E., Jongeneelen, M., Throsby, M., et al. (2009). Antibody recognition of a highly conserved influenza virus epitope. *Science* 324, 246–251.
- Garcia, K., Degano, M., Stanfield, R., Brunmark, A., Jackson, M., Peterson, P., et al. (1996). An alpha beta T cell receptor structure at 2.5 angstrom and its orientation in the TCR-MHC complex. *Science* 274, 209–219.
- Goossens, T., Bräuninger, A., Klein, U., Küppers, R., and Rajewsky, K. (2001). Receptor revision plays no major role in shaping the receptor repertoire of human memory B cells after the onset of somatic hypermutation. *Eur. J. Immunol.* 31, 3638–3648.
- Goossens, T., Klein, U., and Küppers, R. (1998). Frequent occurrence of deletions and duplications during somatic hypermutation: implications for oncogene translocations and heavy chain disease. *Proc. Natl. Acad. Sci. U.S.A.* 95, 2463–2468.
- Han, S., Dillon, S. R., Zheng, B., Shimoda, M., Schlissel, M. S., and Kelsoe, G. (1997). V(D)J recombinase activity in a subset of germinal center B lymphocytes. *Science* 278, 301–305.
- Hertz, M., Kouskoff, V., Nakamura, T., and Nemazee, D. (1998). V(D)J recombinase induction in splenic B lymphocytes is inhibited by antigen-receptor signalling. *Nature* 394, 292–295.
- Hesse, J., Lieber, M., Mizuuchi, K., and Gellert, M. (1989). V(D)J recombination – a functional definition of the joining signals. *Genes Dev.* 3, 1053–1061.
- Hikida, M., Mori, M., Takai, T., Tomochika, K., Hamatani, K., and Ohmori, H. (1996). Reexpression of RAG-1 and RAG-2 genes in activated mature mouse B cells. *Science* 274, 2092–2094.
- Itoh, K., Meffre, E., Albesiano, E., Farber, A., Dines, D., Stein, P., et al. (2000). Immunoglobulin heavy chain variable region gene replacement As a mechanism for receptor revision in rheumatoid arthritis synovial tissue B lymphocytes. *J. Exp. Med.* 192, 1151–1164.
- Jones, J. M., and Gellert, M. (2002). Ordered assembly of the V(D)J synaptic complex ensures accurate recombination. *EMBO J.* 21, 4162–4171.
- Kabat, E. A., Wu, T. T., Gottesman, K. S., and Foeller, C. (1992). *Sequences of Proteins of Immunological Interest*. Darby: Diane Books Publishing Company.
- Kelsoe, G. (1994). B cell diversification and differentiation in the periphery. *J. Exp. Med.* 180, 5–6.
- Kepler, T. B. (1997). Codon bias and plasticity in immunoglobulins. *Mol. Biol. Evol.* 14, 637–643.
- Kiyoi, H., Naoe, T., Horibe, K., and Ohno, R. (1992). Characterization of the immunoglobulin heavy chain complementarity determining region (CDR)-III sequences from human B cell precursor acute lymphoblastic leukemia cells. *J. Clin. Invest.* 89, 739–746.
- Kleinfield, R., Hardy, R. R., Tarlinton, D., Dangl, J., Herzenberg, L. A., and Weigert, M. (1986). Recombination between an expressed immunoglobulin heavy-chain gene and a germline variable gene segment in a Ly 1+ B-cell lymphoma. *Nature* 322, 843–846.
- Kleinfield, R. W., and Weigert, M. G. (1989). Analysis of VH gene replacement events in a B cell lymphoma. *J. Immunol.* 142, 4475–4482.
- Klonowski, K. D., and Monestier, M. (2000). Heavy chain revision in MRL mice: a potential mechanism for the development of autoreactive B cell precursors. *J. Immunol.* 165, 4487–4493.
- Koralov, S. B., Novobrantseva, T. I., Hochedlinger, K., Jaenisch, R., and Rajewsky, K. (2005). Direct *in vivo* VH to JH rearrangement violating the 12/23 rule. *J. Exp. Med.* 201, 341–348.
- Koralov, S. B., Novobrantseva, T. I., Königsmann, J., Ehlich, A., and Rajewsky, K. (2006). Antibody repertoires generated by VH replacement and direct VH to JH joining. *Immunity* 25, 43–53.
- Krause, J. C., Ekiert, D. C., Tumpey, T. M., Smith, P. B., Wilson, I. A., and Crowe, J. E. (2011). An insertion mutation that distorts antibody binding site architecture enhances function of a human antibody. *MBio* 2, e00345–10.
- Kurosawa, Y., and Tonegawa, S. (1982). Organization, structure, and assembly of immunoglobulin heavy chain diversity DNA segments. *J. Exp. Med.* 155, 201–218.

- Lee, A. I., Fugmann, S. D., Cowell, L. G., Ptaszek, L. M., Kelsoe, G., and Schatz, D. G. (2003). A functional analysis of the spacer of V(D)J recombination signal sequences. *PLoS Biol.* 1:e1. doi: 10.1371/journal.pbio.0000001
- Lenze, D., Greiner, A., Knörr, C., Anagnostopoulos, I., Stein, H., and Hummel, M. (2003). Receptor revision of immunoglobulin heavy chain genes in human MALT lymphomas. *Mol. Pathol.* 56, 249–255.
- Lewis, S. M. (1994). The mechanism of V(D)J joining: lessons from molecular, immunological, and comparative analyses. *Adv. Immunol.* 56, 27–150.
- Li, H., Llera, A., and Mariuzza, R. (1998). Structure-function studies of T-cell receptor superantigen interactions. *Immunol. Rev.* 163, 177–186.
- Li, Z., Woo, C. J., Iglesias-Ussel, M. D., Ronai, D., and Scharff, M. D. (2004). The generation of antibody diversity through somatic hypermutation and class switch recombination. *Genes Dev.* 18, 1–11.
- Lutz, J., Müller, W., and Jäck, H.-M. (2006). VH replacement rescues progenitor B cells with two nonproductive VDJ alleles. *J. Immunol.* 177, 7007–7014.
- MacLennan, I. C. (1994). Germinal centers. *Annu. Rev. Immunol.* 12, 117–139.
- MacLennan, I. C., Liu, Y. J., and Johnson, G. D. (1992). Maturation and dispersal of B-cell clones during T cell-dependent antibody responses. *Immunol. Rev.* 126, 143–161.
- Mahajan, K. N., Gangi-Peterson, L., Sorscher, D. H., Wang, J., Gathy, K. N., Mahajan, N. P., et al. (1999). Association of terminal deoxynucleotidyl transferase with Ku. *Proc. Natl. Acad. Sci. U.S.A.* 96, 13926–13931.
- Mansilla-Soto, J., and Cortes, P. (2003). VDJ Recombination: Artemis and its in vivo role in hairpin opening. *J. Exp. Med.* 197, 543–547.
- McBlane, J. F., van Gent, D. C., Ramsden, D. A., Romeo, C., Cuomo, C. A., Gellert, M., et al. (1995). Cleavage at a V(D)J recombination signal requires only RAG1 and RAG2 proteins and occurs in two steps. *Cell* 83, 387–395.
- Montalbano, A., Ogwaro, K. M., Tang, A., Matthews, A. G. W., Larjani, M., Oettinger, M. A., et al. (2003). V(D)J recombination frequencies can be profoundly affected by changes in the spacer sequence. *J. Immunol.* 171, 5296–5304.
- Motoyama, N., Okada, H., and Azuma, T. (1991). Somatic mutation in constant regions of mouse lambda 1 light chains. *Proc. Natl. Acad. Sci. U.S.A.* 88, 7933–7937.
- Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* 102, 553–563.
- Muramatsu, M., Sankaranand, V. S., Anant, S., Sugai, M., Kinoshita, K., Davidson, N. O., et al. (1999). Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. *J. Biol. Chem.* 274, 18470–18476.
- Nadel, B., Tang, A., Lugo, G., Love, V., Escuro, G., and Feeney, A. J. (1998). Decreased frequency of rearrangement due to the synergistic effect of nucleotide changes in the heptamer and nonamer of the recombination signal sequence of the V kappa gene A2b, which is associated with increased susceptibility of Navajos to Haemophilus influenzae type b disease. *J. Immunol.* 161, 6068–6073.
- Nemazee, D. (2006). Receptor editing in lymphocyte development and central tolerance. *Nat. Rev. Immunol.* 6, 728–740.
- Nemazee, D., and Weigert, M. (2000). Revising B cell receptors. *J. Exp. Med.* 191, 1813–1817.
- Nussenzweig, M. C. (1998). Immune receptor editing: revise and select. *Cell* 95, 875–878.
- Oettinger, M., Schatz, D., Gorka, C., and Baltimore, D. (1990). RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. *Science* 248, 1517–1523.
- Papavasiliou, F., Casellas, R., Suh, H., Qin, X. F., Besmer, E., Pelanda, R., et al. (1997). V(D)J recombination in mature B cells: a mechanism for altering antibody responses. *Science* 278, 298–301.
- Pejchal, R., Doores, K. J., Walker, L. M., Khayat, R., Huang, P.-S., Wang, S.-K., et al. (2011). A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. *Science* 334, 1097–1103.
- Pham, P., Bransteitter, R., Petruska, J., and Goodman, M. F. (2003). Progressive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* 424, 103–107.
- Phung, Q. H., Winter, D. B., Cranston, A., Tarone, R. E., Bohr, V. A., Fishel, R., et al. (1998). Increased hypermutation at G and C nucleotides in immunoglobulin variable genes from mice deficient in the MSH2 mismatch repair protein. *J. Exp. Med.* 187, 1745–1751.
- Potter, K. N., Li, Y., and Capra, J. D. (1996). Staphylococcal protein A simultaneously interacts with framework region 1, complementarity-determining region 2, and framework region 3 on human VH3-encoded IgGs. *J. Immunol.* 157, 2982–2988.
- Prabakaran, P., Chen, W., Sengarayan, M. G., Stewart, C. C., Streaker, E., Feng, Y., et al. (2011). Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. *Immunogenetics* 64, 337–350.
- Prak, E. L., and Weigert, M. (1995). Light chain replacement: a new model for antibody gene rearrangement. *J. Exp. Med.* 182, 541–548.
- Raaphorst, F. M., Raman, C. S., Tami, J., Fischbach, M., and Sanz, I. (1997). Human Ig heavy chain CDR3 regions in adult bone marrow pre-B cells display an adult phenotype of diversity: evidence for structural selection of DH amino acid sequences. *Int. Immunol.* 9, 1503–1515.
- Rada, C., Ehrenstein, M. R., Neuberger, M. S., and Milstein, C. (1998). Hot spot focusing of somatic hypermutation in MSH2-deficient mice suggests two stages of mutational targeting. *Immunity* 9, 135–141.
- Rada, C., and Milstein, C. (2001). The intrinsic hypermutability of antibody heavy and light chain genes decays exponentially. *EMBO J.* 20, 4570–4576.
- Radic, M., and Zouali, M. (1996). Receptor editing, immune diversification, and self-tolerance. *Immunity* 5, 505–511.
- Rajewsky, K., Förster, I., and Cumano, A. (1987). Evolutionary and somatic selection of the antibody repertoire in the mouse. *Science* 238, 1088–1094.
- Ramsden, D. A., Baetz, K., and Wu, G. E. (1994). Conservation of sequence in recombination signal sequence spacers. *Nucleic Acids Res.* 22, 1785–1796.
- Ramsden, D. A., McBlane, J. F., van Gent, D. C., and Gellert, M. (1996). Distinct DNA sequence and structure requirements for the two steps of V(D)J recombination signal cleavage. *EMBO J.* 15, 3197–3206.
- Reason, D. C., and Zhou, J. (2006). Codon insertion and deletion functions as a somatic diversification mechanism in human antibody repertoires. *Biol. Direct* 1, 24.
- Reth, M., Gehrmann, P., Petrac, E., and Wiese, P. (1986). A novel VH to VHDJH joining mechanism in heavy-chain-negative (null) pre-B cells results in heavy-chain production. *Nature* 322, 840–842.
- Retter, M. W., and Nemazee, D. (1998). Receptor editing occurs frequently during normal B cell development. *J. Exp. Med.* 188, 1231–1238.
- Roth, D. B. (2003). Restraining the V(D)J recombinase. *Nat. Rev. Immunol.* 3, 656–666.
- Roth, D. B., Menetski, J. P., Nakajima, P. B., Bosma, M. J., and Gellert, M. (1992). V(D)J recombination: broken DNA molecules with covalently sealed (hairpin) coding ends in scid mouse thymocytes. *Cell* 70, 983–991.
- Sadofsky, M. J. (2001). The RAG proteins in V(D)J recombination: more than just a nuclease. *Nucleic Acids Res.* 29, 1399–1409.
- Sanz, I. (1991). Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions. *J. Immunol.* 147, 1720–1729.
- Schatz, D. G. (2004). V(D)J recombination. *Immunol. Rev.* 200, 5–11.
- Schatz, D. G., Oettinger, M. A., and Baltimore, D. (1989). The V(D)J recombination activating gene, RAG-1. *Cell* 59, 1035–1048.
- Schatz, D. G., Oettinger, M. A., and Schlissel, M. S. (1992). V(D)J recombination: molecular biology and regulation. *Annu. Rev. Immunol.* 10, 359–383.
- Schlissel, M., Constantinescu, A., Morrow, T., Baxter, M., and Peng, A. (1993). Double-strand signal sequence breaks in V(D)J recombination are blunt, 5'-phosphorylated, RAG-dependent, and cell cycle regulated. *Genes Dev.* 7, 2520–2532.
- Shapiro, G., and Wysocki, L. (2002). DNA target motifs of somatic mutagenesis in antibody genes. *Crit. Rev. Immunol.* 22, 183–200.
- Shockett, P. E., and Schatz, D. G. (1999). DNA hairpin opening mediated by the RAG1 and RAG2 proteins. *Mol. Cell. Biol.* 19, 4159–4166.
- Steen, S., Gomelsky, L., and Roth, D. (1996). The 12/23 rule is enforced at the cleavage step of V(D)J recombination *in vivo*. *Genes Cells* 1, 543–553.
- Taki, S., Meiering, M., and Rajewsky, K. (1993). Targeted insertion of a variable region gene into the immunoglobulin heavy chain locus. *Science* 262, 1268–1271.
- Throsby, M., van den Brink, E., Jongeneelen, M., Poon, L. L. M., Alard, P., Cornelissen, L., et al. (2008). Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM+ memory B cells. *PLoS ONE* 3:e3942. doi: 10.1371/journal.pone.0003942

- Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature* 302, 575–581.
- van Gent, D. C., Ramsden, D. A., and Gellert, M. (1996). The RAG1 and RAG2 proteins establish the 12/23 rule in V(D)J recombination. *Cell* 85, 107–113.
- Wagner, S. D., Milstein, C., and Neuberger, M. S. (1995). Codon bias targets mutation. *Nature* 376, 732.
- Walker, J. R., Corpina, R. A., and Goldberg, J. (2001). Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. *Nature* 412, 607–614.
- Walker, L. M., Huber, M., Doores, K. J., Falkowska, E., Pejchal, R., Julien, J.-P., et al. (2011). Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* 477, 466–470.
- Walker, L. M., Phogat, S. K., Chan-Hui, P.-Y., Wagner, D., Phung, P., Goss, J. L., et al. (2009). Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* 326, 285–289.
- Watson, L. C., Moffatt-Blue, C. S., McDonald, R. Z., Komppner, E., Ait-Azzouzene, D., Nemazee, D., et al. (2006). Paucity of V-D-D-J rearrangements and VH replacement events in lupus prone and nonautoimmune TdT−/− and TdT+/+ mice. *J. Immunol.* 177, 1120–1128.
- Wiesendanger, M., Kneitz, B., Edelmann, W., and Scharff, M. D. (2000). Somatic hypermutation in MutS homologue (MSH)3-, MSH6-, and MSH3/MSH6-deficient mice reveals a role for the MSH2–MSH6 heterodimer in modulating the base substitution pattern. *J. Exp. Med.* 191, 579–584.
- Wilson, P. C., de Bouteiller, O., Liu, Y., Potter, K., Banchereau, J., Capra, J. D., et al. (1998a). Somatic hypermutation introduces insertions and deletions into immunoglobulin genes. *J. Exp. Med.* 187, 59–70.
- Wilson, P. C., Liu, Y. J., Banchereau, J., Capra, J. D., and Pascual, V. (1998b). Amino acid insertions and deletions contribute to diversify the human Ig repertoire. *Immunol. Rev.* 162, 143–151.
- Wilson, P. C., Wilson, K., Liu, Y. J., Banchereau, J., Pascual, V., and Capra, J. D. (2000). Receptor revision of immunoglobulin heavy chain variable region genes in normal human B lymphocytes. *J. Exp. Med.* 191, 1881–1894.
- Wu, X., Yang, Z.-Y., Li, Y., Hogerkorp, C.-M., Schief, W. R., Seaman, M. S., et al. (2010a). Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* 329, 856–861.
- Wu, Y.-C., Kipling, D., Leong, H. S., Martin, V., Ademokun, A. A., and Dunn-Walters, D. K. (2010b). High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* 116, 1070–1078.
- Zhang, Z., Burrows, P. D., and Cooper, M. D. (2004). The molecular basis and biological significance of VH replacement. *Immunol. Rev.* 197, 231–242.
- Zhang, Z., Zemlin, M., Wang, Y.-H., Munfus, D., Huye, L. E., Findley, H. W., et al. (2003). Contribution of Vh gene replacement to the primary B cell repertoire. *Immunity* 19, 21–31.
- Zheng, N.-Y., Wilson, K., Jared, M., and Wilson, P. C. (2005). Intricate targeting of immunoglobulin somatic hypermutation maximizes the efficiency of affinity maturation. *J. Exp. Med.* 201, 1467–1478.
- Zhou, J., Lottenbach, K. R., Barenkamp, S. J., and Reason, D. C. (2004). Somatic hypermutation and diverse immunoglobulin gene usage in the human antibody response to the capsular polysaccharide of *Streptococcus pneumoniae* Type 6B. *Infect. Immun.* 72, 3505–3514.
- Zhou, T., Georgiev, I., Wu, X., Yang, Z.-Y., Dai, K., Finzi, A., et al. (2010). Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* 329, 811–817.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 20 June 2012; **accepted:** 05 February 2013; **published online:** 11 March 2013.

Citation: Briney BS and Crowe JE Jr. (2013) Secondary mechanisms of diversification in the human antibody repertoire. *Front. Immunol.* 4:42. doi: 10.3389/fimmu.2013.00042

This article was submitted to Frontiers in B Cell Biology, a specialty of Frontiers in Immunology.

Copyright © 2013 Briney and Crowe Jr. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



Age-related changes in human peripheral blood *IGH* repertoire following vaccination

Yu-Chang Bryan Wu¹, David Kipling² and Deborah K. Dunn-Walters^{1*}

¹ Department of Immunobiology, King's College London School of Medicine, London, UK

² Institute of Cancer and Genetics, School of Medicine, Cardiff University, Cardiff, UK

Edited by:

Harry W. Schroeder, University of Alabama, USA

Reviewed by:

Kishore Alugupalli, Thomas Jefferson University, USA

Harry W. Schroeder, University of Alabama, USA

Kay L. Medina, Mayo Clinic, USA

***Correspondence:**

Deborah K. Dunn-Walters, Reader in Immunology, Department of Immunobiology, King's College London School of Medicine, Guy's Campus, London SE1 9RT, UK.

e-mail: deborah.dunn-walters@kcl.ac.uk

Immune protection against pulmonary infections, such as seasonal flu and invasive pneumonia, is severely attenuated with age, and vaccination regimes for the elderly people often fail to elicit effective immune response. We have previously shown that influenza and pneumococcal vaccine responses in the older population are significantly impaired in terms of serum antibody production, and have shown repertoire differences by CDR-H3 spectratype analysis. Here we report a detailed analysis of the B cell repertoire in response to vaccine, including a breakdown of sequences by class and subclass. Clustering analysis of high-throughput sequencing data enables us to visualize the response in terms of expansions of clonotypes, changes in CDR-H3 characteristics, and somatic hypermutation as well as identifying the commonly used *IGH* genes. We have highlighted a number of significant age-related changes in the B cell repertoire. Interestingly, in light of the fact that IgG is the most prevalent serum antibody and the most widely used as a correlate of protection, the most striking age-related differences are in the IgA response, with defects also seen in the IgM repertoire. In addition there is a skewing toward IgG2 in the IgG sequences of the older samples at all time points. This analysis illustrates the importance of antibody classes other than IgG and has highlighted a number of areas for future consideration in vaccine studies of the elderly.

Keywords: B cell repertoire, aging, vaccine, immunoglobulin, IgA

INTRODUCTION

Until recently the methods available to study B cell repertoire were limited by the fact that the diversity of the repertoire was far greater than the number of sequences that could feasibly be studied. A random sampling of cells responding to vaccine would pick up the most prevalent Ig genes but would not give any indication of the diversity of cells responding to challenge. In mice the T-dependent NP response is a widely used and versatile tool for the study of immune responses, it tends to be restricted to use of the V186.2 heavy chain. In humans the isolation of rearranged Ig genes with more than one type of *IGHV* gene has indicated that there might be diversity in the response, but the numbers of sequences studied in these experiments have been low. One of the best examples of repertoire analysis was by Kolibab et al. (2005a) who looked at approximately 1300 sequences from 40 different donors after immunization by the pneumococcal vaccine and identified the major *IGHV* genes in use. With the advent of high-throughput sequencing methods we can now study the human immune response in much more detail.

A perennial problem in vaccination is that of vaccine inefficiency in older people. Vaccine-specific antibodies in older people are quantitatively and qualitatively impaired. For example, anti-pneumococcal polysaccharide (PPS) antibodies have lower opsonophagocytic index and affinity in the elderly than in healthy young adults (Kolibab et al., 2005b; Park and Nahm, 2011). *Streptococcus pneumoniae* infection is a serious complication secondary

to influenza, and together these two diseases comprise a substantial infectious burden for children aged under two, the elderly and immunocompromised individuals (McCullers, 2006). In the UK and many other countries, co-administration of trivalent-influenza and pneumococcal vaccines has been recommended as the routine immunization schedule to protect these at-risk groups. Although influenza vaccines appear effective in most groups with nearly 70% vaccine efficacy (Osterholm et al., 2012), less than half of the older adults are protected by influenza vaccines (Nichol et al., 2007). While the immunogenicity of PPS vaccines in the protein-conjugated form (PCV) is much improved for young children, neither PCV nor 23-valent PPS vaccine (PPV-23) is able to confer an effective protection in the older population (Baxendale et al., 2010).

We have previously seen that there may be changes in the selection process during affinity maturation of B cells (Banerjee et al., 2002), and also that the B cell repertoire is often less diverse in old age with evidence of non-pathogenic clonal expansions (Gibson et al., 2009). This loss of diversity correlated with the health of the individual. Further investigation as to whether loss of diversity in the B cell repertoire correlated with poor vaccine responses against influenza and pneumonia indicated that it may be a contributory factor, but that other factors were likely also involved (Ademokun et al., 2011). These two vaccines generate different responses; influenza is believed to mainly induce IgG1/IgG3/IgA1 T-dependent responses (Brown et al., 1985; Hocart et al., 1990;

Powers, 1994), while pneumococcal responses are thought to be T-independent and can lead to significant elevation of IgA2/IgG2 antibodies in the serum and mucosa (Lue et al., 1988; Carson et al., 1995; Sanal et al., 1999; Simell et al., 2006; Benkert et al., 2011).

To investigate the diversity of the response to these vaccines in more detail we have analyzed high-throughput sequencing data in order to characterize the response with respect to Ig gene usage and hypermutation whilst paying particular attention to the subclasses of antibodies involved. There are significant differences in the older vaccine response with respect to class and subclass of antibody, extent and timing of clonal expansions and focusing of the repertoire toward Ig sequences with higher mutation and shorter CDR-H3 regions.

MATERIALS AND METHODS

VOLUNTEERS AND SAMPLE COLLECTION

Six young (aged 19–45) and six older (aged 70–89) healthy volunteers were recruited as a part of the 2009/10 influenza vaccination program in Lambeth Walk GP Practice. Blood and serum samples were collected after obtaining written consent as approved by the Guy's Hospital Research ethics committee, prior to vaccination at day 0 (D0) with the influenza (Influvac; Solvay, Southampton, UK), and 23-valent pneumococcal (Pneumovax II; Sanofi Pasteur MSD, Maidenhead, UK) vaccines and post vaccination at day 7 (D7) and day 28 (D28). PBMCs were isolated using Ficoll-Paque Plus (GE Healthcare, Buckinghamshire, UK) in conjunction with Leucosep tubes (Greiner Bio-One Ltd., Gloucestershire, UK), according to manufacturer's instructions.

TOTAL RNA EXTRACTION AND cDNA CONVERSION

Total RNA was extracted from 2×10^7 PBMCs per donor per time point using the RNeasy Mini Kit (Qiagen, UK). The SuperScript III First-Strand cDNA Synthesis System (Invitrogen, UK) was then used to convert RNA into cDNA according to the manufacturer's protocol. In brief, a 200- μ L cDNA reaction mix contained extracted total RNA, 500 ng Oligo(dT)₂₀, 500 μ M dNTPs, 400 U RNaseOut, 10 mM DTT, and 2000 U SuperScript III RT in 1 \times First-Strand RT buffer. The cDNA reaction was carried out as follows: 65°C for (5 min), 4°C (60 s); 50°C (1 h); 70°C (15 min).

HIGH-THROUGHPUT SEQUENCING OF *IgH* GENES

Ig genes were isolated by semi-nested, isotype-specific PCR reactions, as previously reported (Wu et al., 2010). Briefly, a 25- μ L PCR1 reaction mix contained 6.25 μ L of cDNA, 0.625 U Phusion DNA polymerase (NEB, UK), 200 μ M each dNTP, 41.75 nM each of 6 *IGHV* gene family primers in conjunction with 250 nM of either IgM, pan-IgA, or pan-IgG constant region primers. Two microliters of PCR1 products were subsequently re-amplified using multiplex identifier (MID)-containing primers in a semi-nested reaction consisting of 0.5 U Phusion DNA polymerase, 200 μ M each dNTPs, 41.75 nM each of the *IGHV* gene family/MID primers, and 250 nM of the nested constant region/MID primers in a 20- μ L reaction volume. PCR conditions are as follows: 98°C for (30 s), 15 (PCR1), or 20 (PCR2) cycles of 98°C (10 s); 58°C (15 s); 72°C (45 s), and 1 cycle of 72°C (10 min). In order to produce sufficient quantity for high-throughput sequencing on the GS FLX System (Roche, Germany), eight different PCR1 for

each sample, followed by 16 PCR2 (two per initial PCR1 round) reactions were performed for each isotype. This total sampling of cDNA was approximately equivalent to that from 2×10^5 B cells. The downstream preparation of PCR products and data processing are as previously published (Wu et al., 2010).

DETECTION OF ANTI-PPS IgA ANTIBODIES

Serum anti-PPS IgA antibodies from immunized young (age 18–49, $n = 39$) and older (age 65–89, $n = 27$) healthy volunteers were measured by ELISA, as previous reported (Ademokun et al., 2011). In brief, the 89-SF standard (Bethesda, MD, USA) or serum samples were pre-absorbed with 10 μ g/mL cell wall polysaccharide (CPS; Statens Serum Institute, Copenhagen, Denmark) or 10 μ g/mL CPS in conjunction with 10 μ g/mL 22F polysaccharide, respectively, before being serial diluted and incubated with microtiter plates coated with 100 μ L per well of a combination of seven polysaccharide serotypes (4, 6B, 9, 14, 18C, 19F, and 23F, 1 μ g/mL each; ATCC, Rockville, MD, USA) overnight at 4°C. After washing with TBS containing 0.1% Brij 35 (Sigma-Aldrich, St. Louis, MO, USA), HRP-conjugated goat anti-human IgA (Invitrogen) diluted at 1/4000 in PBS with 0.05% Tween-20 was added to the plates and incubated for 2 h, followed by another 2 h of incubation with TMB chromogen substrate (Invitrogen) in diethanolamine buffer, before terminating the reaction with 3 M NaOH. Serum anti-PPS IgA levels were read at OD 450 nM on an ELISA microplate reader and then compared with the 89-SF standard.

SEQUENCE ANALYSIS

Ig gene usage and CDR-H3 junction regions between the conserved first (cysteine) and last amino acid (tryptophan) were determined using IMGT V-QUEST (Wilkins et al., 1999). Internal isotype motifs in the constant regions of each sequence further identified subclasses of sequences. ProtParam was used to determine the physicochemical properties of the CDR-H3 peptide (Brochet et al., 2008). The grand average of hydropathicity (GRAVY) and aliphatic indices are positive indicators for peptide hydrophobicity and structural thermostability respectively (Ikai, 1980; Wilkins et al., 1999). The percentage match of each *IGHV* sequence to germline gene was returned by IMGT V-QUEST, and the level of hypermutation was calculated to be the percentage difference from the corresponding germline gene.

The DNA sequences of the CDR-H3 were used for clonotype clustering by a distance-matrix between all pairwise comparisons, as previously reported (Ademokun et al., 2011). When necessary, clonally related IGH sequences were aligned with putative germline genes and edited to remove homopolymer tract errors using DNASTar software (Laser Gene). Mutational phylogenetic trees of the edited IGH sequences were constructed by the multiple sequence alignment modes (MUSCLE 3.7) using the Phylogeny Analysis program (Dereeper et al., 2010).

STATISTICS

Statistics were performed with GraphPad Prism 5.0. Most statistical analyses were one-way or repeated measures ANOVA (with Bonferroni post-test) and Kruskal-Wallis comparisons (with Dunn's post-test). Wherever necessary, Chi-squared test (with

Bonferroni post-test), paired *t*-test, and Mann–Whitney *U*-test was performed. To test association between different metrics, Pearson correlation analysis in conjunction with linear regression was performed.

RESULTS

IGH REPERTOIRE CHANGES IN RESPONSE TO VACCINATION

High-throughput sequencing of samples from peripheral blood B cells in six young and six old donors in the course of vaccination produced 45,784 *IGH* sequences, which were grouped into 17,962 different clonotypes (i.e., a representative clone for that particular Ig gene rearrangement, **Table 1**). In order to investigate the effect of vaccination on the underlying repertoire without the influence of *in vivo* and *in vitro* clonal expansion, only clonotypes were included for this analysis. In general, the repertoire displayed resilience in that at D28 post vaccination it showed similar characteristics to D0 (before vaccination) despite significant changes at day 7 post vaccination. These changes at D7 included a change in expression of some *IGHV* and *IGHJ* genes such that significant differences were seen in *IGH* gene family usage (**Figure 1A**). Most notably, at D7 after vaccination there is a significant increase in the proportion of clonotypes that use *IGHV6-1*, *IGHV1-46*, and several *IGHV3* genes with a decrease in those using *IGHV2*, *IGHV4* gene families, and *IGHV3-21* (**Figure 1B**). An increased usage of the shorter *IGHJ4* genes by 4.5% with a concomitant decrease in use of the longer *IGHJ6* gene (**Figure 1C**; $p < 0.05$) at D7 may explain the overall reduction in the CDR-H3 size by 2 nt (**Figure 1D**; $p < 0.0005$, Kruskal–Wallis test). The D7 population also has a more hydrophilic CDR-H3, as shown by a more negative GRAVY index (**Figure 1E**; $p < 0.0005$), and is also less aliphatic (**Figure 1F**; $p < 0.0005$).

AGE-RELATED DIFFERENCES IN VACCINE-INDUCED CLONAL EXPANSION

It has been shown that clonal expansion after challenge is delayed and persists longer in old mice (Szabo et al., 2004). More recently, Lindner et al. (2012) reported that the intestinal IgA repertoire has fewer large, expanded, clones in old mice. In humans, we have previously shown vaccine-induced changes in CDR-H3 size and hydrophobicity in expanded clones of young people, but the changes are less obvious in the older group (Ademokun

et al., 2011). Here we aim to investigate age and challenge-related changes in *IGH* gene usage. To ensure that our comparative analyses of clonal expansion reflect the effects of *in vivo* clonal expansion, rather than *in vitro* amplification from PCR, the *IGH* sequences were produced using the same number of cells in each sample. In the overall repertoire, large clones containing up to 687 member sequences appear at D7 and are seen less frequently at D28 in the young group, whereas in the older group the largest clone that appears at D7 has only 242 sequence members and some large clones are also observed at D28 (**Figure 2A**). Intraclonal variations in *IGHV* sequences indicate that these large clones likely represent *in vivo* expansion (**Figure 6**). To understand how clonal expansion occurs in different classes of B cells, *IGH* sequences were grouped by isotype for further analyses of clone size (**Figures 2B–D**). Significant increases in the average clone size at D7, as result of increased proportion of large clones, are seen in all isotypes in both age groups, although the increase in IgM clones is not as great as that seen in IgA and IgG clones. The pattern of vaccine responses in IgG and IgM clones, as indicated by changes in the average clone size, is similar between the two age groups (**Figure 2B**). The increase in average clone size was usually related to the accumulation of greater numbers of larger clone sizes, except in the case of IgA, where the young and old samples had comparable numbers of clones larger than 3, but the average clone size in young IgA clones at D7 is twice that of the old clones at D7 (**Figure 2B**; $p < 0.005$). This implies that the IgA clones in the young are larger than in the old. Consistent with this, the largest clones at D7 in the young in **Figure 2A** were IgA. Interestingly, the clone size is 2.5-fold larger in old IgA clones at D28 as compared with the young (**Figure 2B**; $p < 0.005$), and is larger than it was in the old at D7. So there may be a more delayed clonal expansion in IgA cells with age; while the young showed expansion at D7 and contraction back to D0 levels at D28, the level of expansion in the old may not yet have reached a peak by D28.

The vaccine response at D7 is very diverse at all ages in that changes in the average clone size are not restricted to particular *IGHV* family *IGHJ* gene combinations (**Figure 2C**). We calculated the fold difference in clone size between D7/D0 and D28/D7 for individual *IGHV* genes. At D7 there were 24 different *IGHV* genes of the IgA isotype and 16 different *IGHV* genes of the IgG isotype that had more than twofold increases in their average clone size

Table 1 | Numbers of *IGH* sequence and clonotypes produced by high-throughput sequencing.

Ages	Days	IgA	IgG	IgM	Unknown ³	Total
Young (6 donors)	D0	886 ¹ /1424 ²	1035/1651	847/1208	413/737	3181/5020
	D7	976/5922	717/2707	641/1214	4912/750	2825/12593
	D28	1070/1777	1298/2085	1036/1280	626/1059	4030/6201
Old (6 donors)	D0	928/1747	675/1528	809/1109	474/1175	2886/5559
	D7	644/1994	988/3801	579/998	418/1825	2629/8618
	D28	1223/5099	475/901	322/400	391/1393	2411/7793

¹The numbers of clonotypes refer to unique sequences, where only one example of a clonal expansion is counted after CDR-H3 clonotype clustering.

²The numbers of *IGH* sequences, generated from approximately 2×10^6 B cells per sample, represent those that passed quality control with full immunoglobulin VDJ gene rearrangement.

³Unknown isotype refers to sequences that do not extend far enough to the constant region for isotype identification.

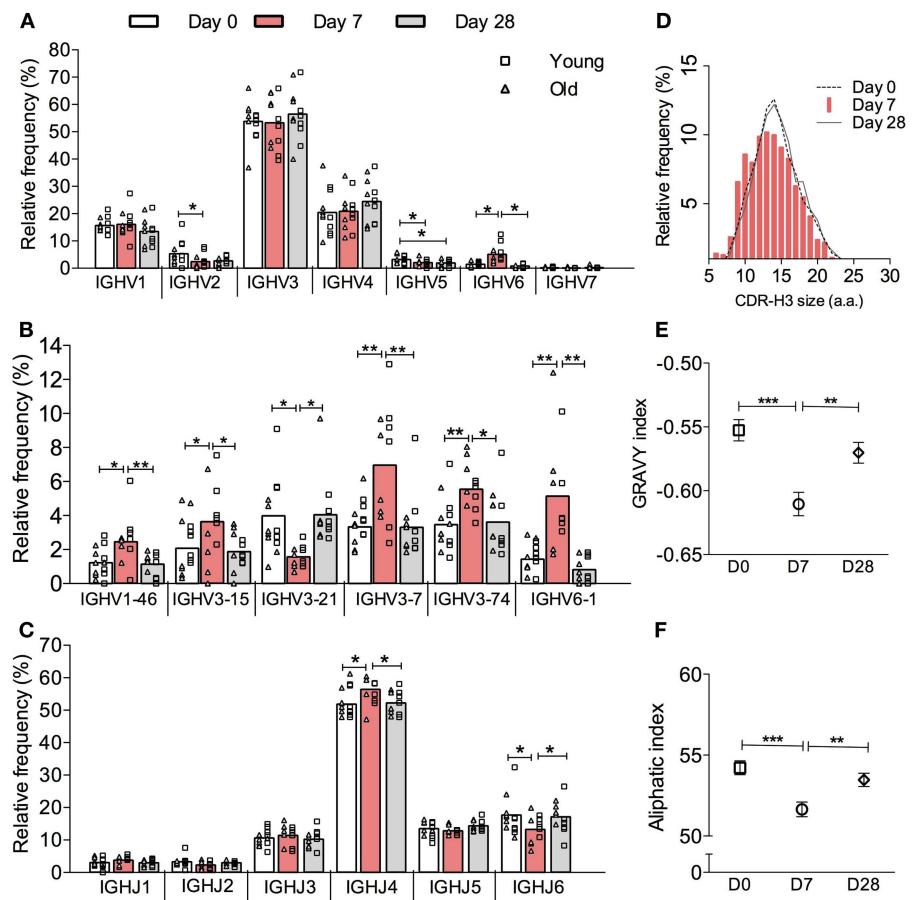


FIGURE 1 | Vaccine-induced changes in the overall *IGH* clonotype repertoire. *IGH* genes in each sequence were analyzed by IMGT V-QUEST and then clustered into clones by CDR-H3 DNA sequence similarity. Only one sequence example per clone was chosen to represent its clonotype. The relative frequency (y-axis) of all clonotypes (IgA, IgG, and IgM combined) by (A) *IGHV* gene family, (B) individual *IGHV* gene, and (C) *IGHJ* gene usage (x-axis) was calculated for six young (squares) and six old (triangles) donors individually before being grouped as a whole cohort of 12 donors for repeated measures ANOVA comparison between D0 (open bars), D7 (red bars) and D28 (grey bars).

Bars indicate MEAN. (D) CDR-H3 virtual spectratypes, showing the relative frequency (y-axis) of a particular CDR-H3 size (x-axis in amino acid numbers) within the whole CDR-H3 repertoire, was calculated using all clonotypes from 12 donors at D0 (dotted line), D7 (red bars), and D28 (solid line). The GRAVY and aliphatic indices for each CDR-H3 peptide was determined by ProtParam. The overall GRAVY (E) and aliphatic (F) indices all clonotypes from 12 donors were compared by Kruskal-Wallis comparisons with Dunn's post-test between D0 (squares), D7 (circles), and D28 (diamonds). Error bars indicate \pm SEM. * $p < 0.05$, ** $p < 0.005$, and *** $p < 0.0005$.

(Figure 2D). The above age-related differences in the average clone size for IgA clones at D7 and D28 (Figure 2B) are a reflection of a diverse response, involving many different *IGHV* family *IGHJ* combinations and individual *IGHV* genes rather than any particular *IGH* genes (Figures 2C,D). The changes in the average clone size and serum anti-PPS IgA antibodies at D7 from D0 were also negatively correlated with age (Figures 2E,F).

AGE-RELATED DIFFERENCES IN CDR-H3 CHARACTERISTICS OF EXPANDED CLONES

CDR-H3 regions have been regarded as indispensable Ig structures for antigen recognition, therefore alterations in the physicochemical properties of CDR-H3 peptides could potentially affect BCR binding ability and selection and expansion of antigen specific cells within a population may be detectable by changes in the CDR-H3 characteristics (Romero-Steiner et al., 1999; Kolibab et al.,

2005b; Park and Nahm, 2011). We therefore compared CDR-H3 characteristics in *IGH* sequences from clones of different sizes at D7 after challenge. The clonotypes were divided into those seen four or less times in the sample (small clones) and those seen five or more times (large clones). The overall hydrophobicity and aliphatic index of CDR-H3 regions from large clones does not significantly differ between ages when all isotypes are considered together (data not shown). However, there are significant age-related changes in CDR-H3 size in IgA and IgM clonotypes from large clones (Figure 3A) that was not seen in IgG clonotypes (data not shown). IgM clonotypes show significantly larger CDR-H3 sizes in all clones regardless of clone size in the old samples than in the young at both D0 and D7 (Figure 3A). The differences in the CDR-H3 size are mainly due to an increase in the total length of *IGHD* genes (Figure 3B), not *IGHJ* (data not shown), and an increase in the numbers of N-nucleotides that reaches significance

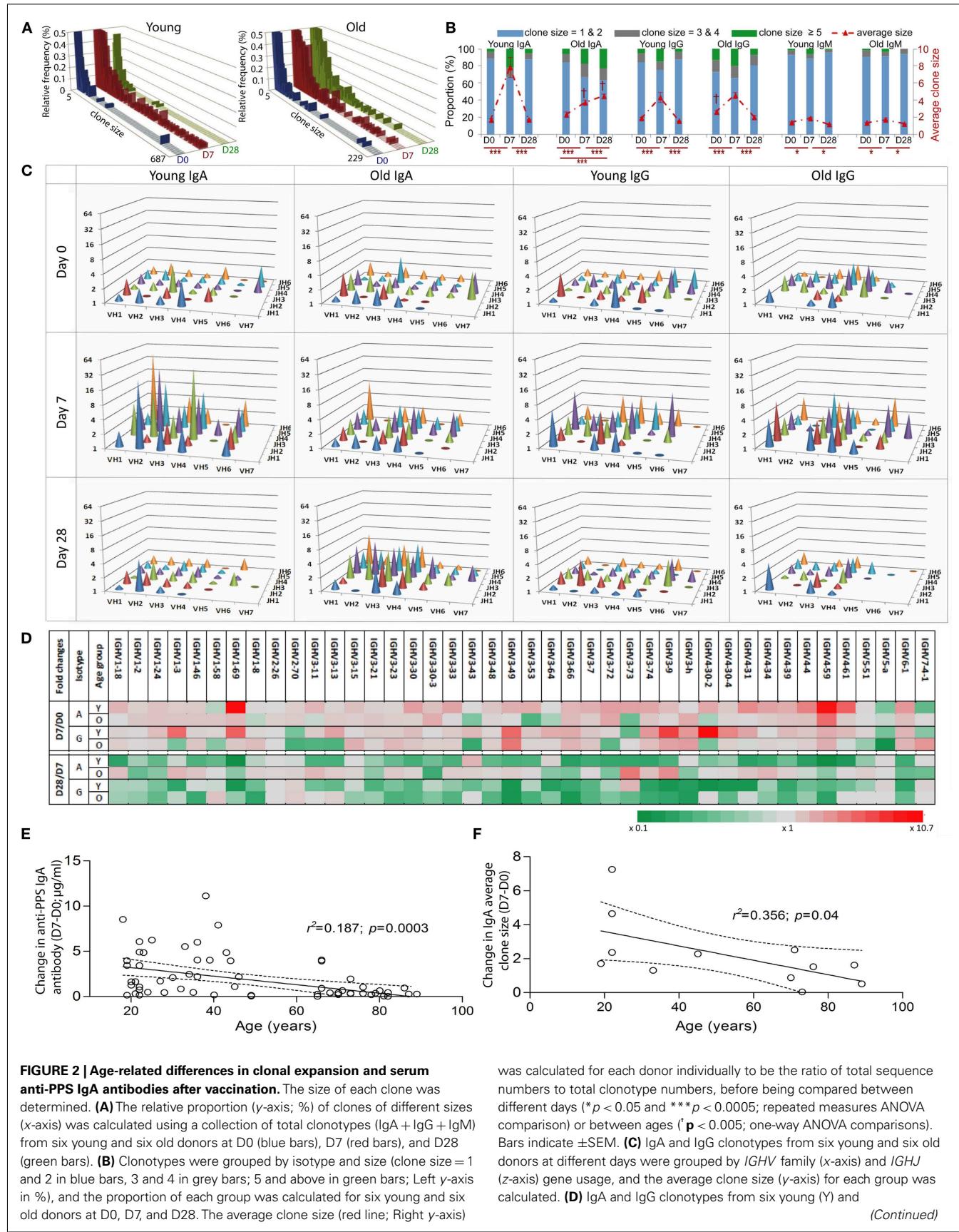


FIGURE 2 | Continued

six old (O) donors at different days were grouped by individual *IGHV* gene usage, and the average clone size for each group was calculated. The fold change of the average clone size at D0 from D7 (D7/D0) and at D28 from D7 (D28/D7) was calculated and is shown by different colors (grey as unchanged, green as fold decrease and red as fold increase). Serum levels of IgA

antibodies specific for seven serotypes combined were determined using ELISA. Pearson correlation analysis was used to test the correlative relationship of age (x-axis in years) with changes in (E) serum anti-PPS IgA antibodies ($n=66$) and (F) the average clone size (D7-D0, $n=12$), r^2 and p values indicated. The Goodness-of-Fit (solid lines) was analyzed by linear regression with 95% CI indicated (dashed lines).

in expanded large clones (Figure 3C). There is also a significant age-related increase in the CDR-H3 size and CDR-H3 components in IgA clonotypes, although this is restricted to D7 and D28 post-challenge and is mainly in the larger clones (Figures 3A–C).

AGE- AND CHALLENGE-RELATED CHANGES IN HYPERMUTATION

Since the level of hypermutation accumulated in the Ig variable region often reflects the history of affinity maturation in B cell clones specific for a particular type of antigen, we compared *IGHV* mutation levels over the course of vaccination and in different age groups. The average mutation frequency is significantly increased in IgM clonotypes from both age groups at D7 after vaccination (Figure 4A), although the change at D7 from D0 in old IgM clones was threefold smaller than that in the young. Changes in overall *IGHV* mutation levels in IgM clones in response to vaccines and with age can also be demonstrated by the proportional alteration between unmutated versus heavily mutated IgM clones (Figures 4B,C).

A higher mutational frequency is observed with age in IgG clonotypes at all time points ($p=0.0003$, Mann–Whitney U -test; data not shown). We did not see an increase in the overall mutation levels in IgA clonotypes after challenge in either age group (data not shown). Since IgA populations are reflective of prior challenge and hypermutation this perhaps would not be expected. However, looking solely at the large clones (five sequences or more per clone), presumably enriched for the ones responding to vaccination as opposed to those that were historically mutated, we saw that IgA clonotypes from old donors had a lower average frequency of mutation than the young ($p=0.0025$, Mann–Whitney U -test). This may be due to the reduction in the proportion of heavily mutated, large IgA clonotypes with age (Figures 4D,E). A similar trend is also observed in IgM clonotypes (Figure 4F).

AGE AND CHALLENGE-RELATED CHANGES IN CLONOTYPES AND CLONAL EXPANSION BY IGH SUBCLASS

We previously reported that isotype-switched and innate-like IgM+ memory cells have distinct *IGHV* repertoires, and later that there are some similarities between the innate-like IgM+ memory cell repertoire and that of IgG2 and (to a lesser extent) IgA2 memory cells. Therefore we proposed that a large proportion of IgG2 and IgM (and possibly IgA2) memory cells may be subject to a different selection process from that imposed on other classes and subclasses of memory cells (Wu et al., 2010, 2011). In order to compare the vaccine response in B cells of different subclasses we stratified the overall clonotype repertoire by subclass. This was done by searching for subclass-specific motifs upstream of the pan-IgA or pan-IgG primer sequences, so that any difference in subclass within a particular class is not due to a difference in primer binding efficiency. In line with previous reports (Wu et al., 2010, 2011), over-representation of *IGHV3* and under-representation of *IGHV1* genes are seen in IgM, IgG2, and IgA2 clonotypes in both age groups at all time points analyzed when compared to IgG1, IgG3, and IgA1 clonotypes (data not shown). Within the IgA and IgG compartments the proportions of the subclasses vary with challenge, with a significant increase in the proportion of IgA2 and IgG2 clonotypes being seen in both age groups at D7 compared to D0 (Figure 5A; $p<0.005$, repeated measures ANOVA). Although we failed to detect any age-related differences in Ig gene usage in IgA and IgG subclasses (data not shown), there is a significant difference in IgG subclass distribution with age at both D0 and D7, with the older group having an increased proportion of IgG2 compared with the young (Figure 5A; $p<0.0005$, Mann–Whitney U -test). The increased proportion of IgA2 and IgG2 subclasses after vaccination are also observed when clone relatives (i.e., all sequences not just clonotypes) are included for analysis (Figure 5B).

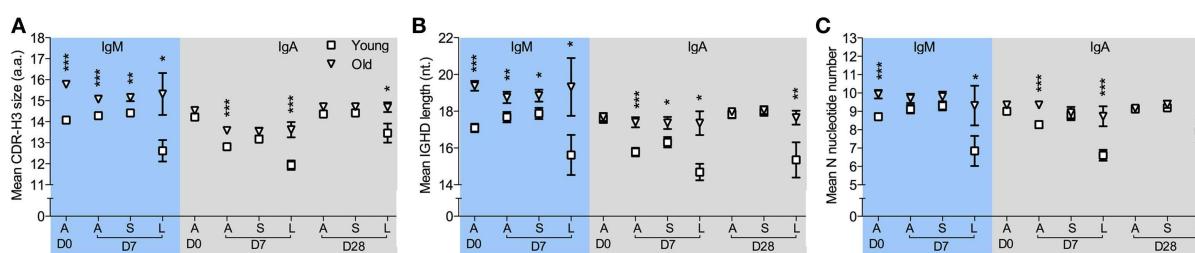


FIGURE 3 | Age-related changes in CDR-H3 characteristics from clonotypes in different sizes. IMGT V-QUEST was used to determine the size of the CDR-H3 region and its components, i.e., N-nucleotide numbers and *IGHD* genes, in each *IGH* sequence. After clustered by CDR-H3 sequence motifs, IgM, and IgA clonotypes were sorted by clone sizes. (A) The mean size of CDR-H3 regions (in amino acids),

(B) the mean length of overall *IGHD* genes (in nucleotides), and (C) the mean number of N-nucleotides from all clonotypes (A), clonotypes ≤ 4 (S) and clonotypes ≥ 5 in size (L) were compared between six young (square) and six old (triangle) donors at D0, D7, and D28, using Mann–Whitney U -tests (* $p<0.05$, ** $p<0.005$, and *** $p<0.0005$). Error bars indicate \pm SEM.

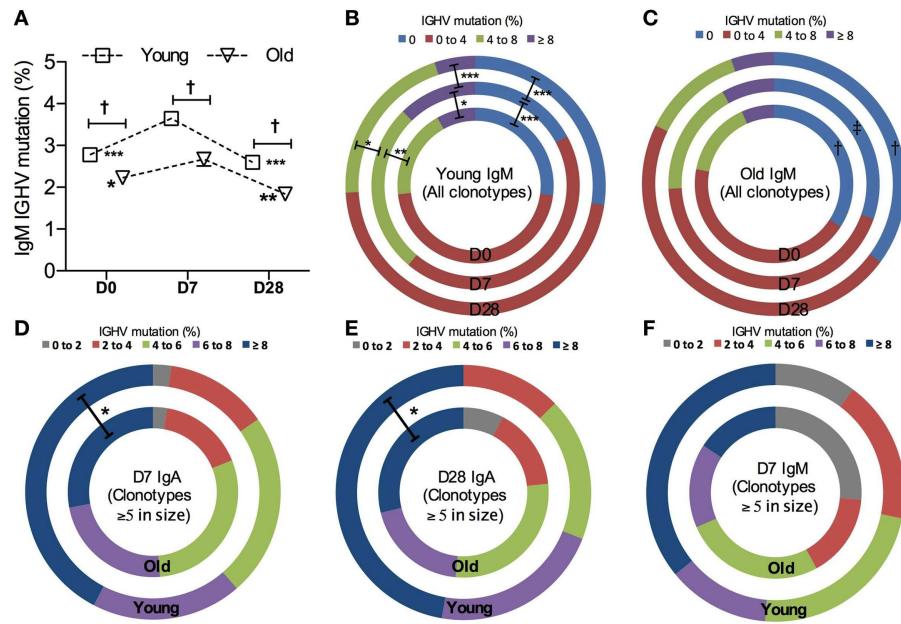


FIGURE 4 | Vaccine-induced changes in IGHV mutations. (A) The frequency of IGHV mutations in each sequence was calculated to be the percentage difference from germline identity, determined by IMGT V-QUEST. Kruskal–Wallis comparisons were used to compare the mean IGHV mutation between IgM clonotypes from six young (squares) and six old (triangle) donors (age-related differences: $^{\dagger}p < 0.0005$) and between D0, D7, and D28 (temporal differences: $^{*}p < 0.05$, $^{**}p < 0.005$, and $^{***}p < 0.0005$, as compared with D7). Error bars indicate \pm SEM. All IgM clonotypes were grouped by the level of IGHV mutation and the

proportion of each group was compared using Chi-squared tests between days (D0: inner circles, D7: middle circles, and D28: outer circles; $*p < 0.05$, $^{**}p < 0.005$, and $^{***}p < 0.0005$) and between six young (B) and six old (C) donors [$^{\dagger}p < 0.05$ and $^{+}p < 0.005$, as indicated in (C)]. For clones ≥ 5 in size, IgM clonotypes at D7 (D), IgA clonotypes at D7 (E), and IgA clonotypes at D28 (F) were grouped by the level of IGHV mutations and the proportion of each group was compared between six young (outer circles) and six old (inner circles) donors, using Chi-squared tests ($^{*}p < 0.05$).

Generally the differences in clonotype subclass distribution (Figure 5A) are mirrored in sequence subclass distribution (Figure 5B). However, in IgG there appears to be a greater bias toward IgG2 in sequences compared to clonotypes, indicating that in the young the IgG2 clonotypes belong to larger clones. Differences in IgA and IgG average clone size at the different time points are shown in Figures 5C,E respectively, with details of the distribution of clones in the different subclasses in Figures 5D,F. Significant age-related differences in the average IgA clone size (Figure 2B) are seen at D7 and D28, and appear to be mainly due to the IgA1 subclass having a smaller clone size at D7 but bigger at D28 in the old as compared with the young (Figure 5C). Although the average clone size in IgA2 does not seem to change with age, more detailed analysis of IgA2 clone size distribution shows that the frequency of large clones, containing over 20 sequence members, is lower at D7 ($p = 0.002$; Chi-squared test) but higher at D28 in old repertoires, as compared with the young (Figure 5E). In both age groups the average clone size of IgG2 is significantly bigger than IgG1 at D7 ($p = 0.0008$, paired t -test; Figure 5D). Thus, in addition to there being an increase in the number of different IgG2 clonotypes in the sampled repertoire at D7, the individual IgG2 clonotypes are expanded more than IgG1 clonotypes. The age-related increase in the proportion of IgG2 clonotypes at D7 and D28 (Figure 5A) is not accompanied by any significant difference in the clone size (Figure 5D). So although the older repertoire has a greater representation of IgG2 sequences these are not expanded

any more than in the younger group. The IgG1 and IgG2 clone distribution is illustrated in more detail in Figure 5F.

CLONAL EXPANSION AS A RESULT OF RECALL IMMUNE RESPONSES

Previous studies show that influenza-specific cells (Wrammert et al., 2008) and anti-PPS plasma cells (Baxendale et al., 2010) can be detected in the serum prior to vaccination and their frequencies are increased at D7 following vaccination. Although our sequences were produced using unsorted PBMCs, large sequence numbers allowed us to track responding clones sampled at D7 back to their clone relatives sampled at D0 prior to vaccination, using mutational phylogeny analysis (Dereeper et al., 2010). We find a total of 648 different clonotypes (3.6% of all clonotypes) containing clonally related IGH sequences that represent cells sampled at different days (Table 2). 71% of these clones are of a single isotype, being significantly more frequent than those having IGHM sequences related to IGHG and IGH α sequences (12.8%; Chi-squared test, $p < 0.0001$). As expected, the proportion of clonotypes that share sequences across different time points increases to 20%, when only large clones are considered (Table 2). Out of these larger clones, 57 clonotypes (4.6%) contain IGH sequences that have already switched to IgA and IgG isotypes at D0 prior to vaccination (Figure 6A), suggesting a recall immune response by pre-existing memory B cells. Similarly, there are also clones that contain mutated IgM+ sequences at D0 (Figure 6B). Mutational phylogenetic trees show great diversification within

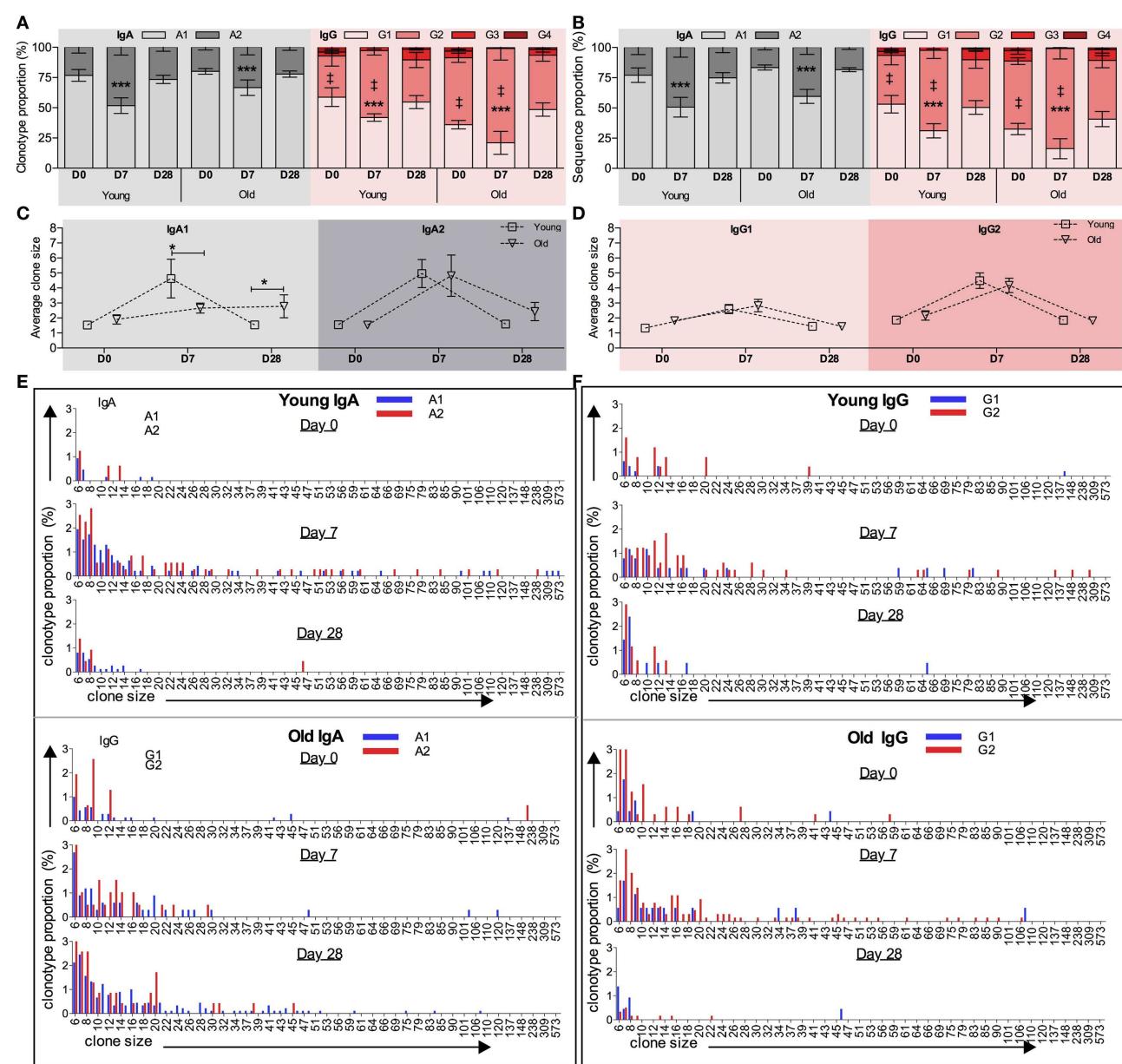


FIGURE 5 | Age and challenge-related changes in IGH subclass.

Internal motifs in the constant region were used to identify subclasses (IgA1, IgA2, IgG1, IgG2, IgG3, and IgG4) in each IGH sequence. The proportion (y-axis) of clonotypes clustered by CDR-H3 and total sequences (y-axis) of different IgA or IgG subclasses at D0, D7, and D28 was calculated for each donor individually before being collectively analyzed (temporal differences: *** $p < 0.0005$ using repeated measures ANOVA when compared with D7; age-related differences: * $p < 0.05$ using Mann–Whitney U-tests). Error bars indicate \pm SEM. The average clone

size (y-axis) of IgA1 and IgA2 clonotypes (**C**) and IgG1 and IgG2 clonotypes (**D**) was calculated to be the ratio of total sequences over total clonotypes for each donor individually, before being compared between young versus old ages (* $p < 0.05$; one-way ANOVA comparison). Bars indicate \pm SEM. Clonotypes of (**E**) IgA subclasses (IgA1: blue bars and IgA2: red bars) and (**F**) IgG subclasses (IgG1: blue bars and IgG2: red bars) were grouped by their clone sizes (x-axis), and the proportion of each groups (y-axis) was calculated at D0, D7, and D28. Clones containing fewer than five sequences are not shown.

expanded clones. Interestingly we also observe that sequences sampled at D0 do not always appear less mutated than those at D7 and D28. Similarly, IGHM sequences do not always appear before IGH α /IGH γ sequences in the lineage tree. These observations have important consequences for future interpretation of data based on analysis of phylogenetic trees.

DISCUSSION

Improving the immunogenicity of vaccines against *S. pneumoniae* and influenza in the older person is a challenge (Artz et al., 2003; Hannoun et al., 2004). Effective antibody production is impaired in older people but the exact causes of this impairment have not been fully elucidated. An earlier study showed that affinity

Table 2 | Numbers of clones with clonality between *IGH* sequences sampled at different days¹.

All clonotypes ²										
Day ¹	Old (n = 7926 ³ ; 6 donors)					Young (n = 9774 ³ ; 6 donors)				
ISO ⁵	AG	AGM	AM	GM	Single ⁶	AG	AGM	AM	GM	Single
0 and 28	3	1	1	1	94	17	0	3	7	94
0 and 7	4	3	7	4	47	8	3	20	1	38
0 and 7 and 28	4	1	2	0	16	5	3	2	0	4
7 and 28	23	4	3	3	93	38	6	8	0	77

Clonotypes ≥5 in size										
Day	Old (n = 671 ⁴ ; 6 donors)					Young (n = 572 ⁴ ; 6 donors)				
ISO	AG	AGM	AM	GM	Single	AG	AGM	AM	GM	Single
0 and 28			1	1	25	11	0	0	1	22
0 and 7	3	3	1	1	19	5	2	13	1	9
0 and 7 and 28	4	1	1	0	12	4	3	1	0	2
7 and 28	20	4	2	0	36	32	6	3	0	24

¹CDR-H3 sequence motifs are used to identify clones containing *IGH* sequences that have the same VDJ rearrangement but are sampled at Day 0, 7, and 28.

²All clonotype refers to all clones in various sizes.

³Numbers refers to all clonotypes produced by high-throughput sequencing.

⁴Numbers refer to clonotypes ≥5 in size produced by high-throughput sequencing.

⁵Iso refers to clones containing *IGH* sequences sharing the same CDR-H3 region but are of different isotypes (A for IgA, G for IgG, and M for IgM).

⁶Single refer to clones containing *IGH* sequences of only one isotype.

selection in the germinal center may be altered with age, this may be partly due to lack of appropriate help, for example from T cells, as well as intrinsic differences in the B cell affecting its ability to express AID (Banerjee et al., 2002; Frasca et al., 2008). We later showed that B cell diversity decreases with age and is associated with poor health (Gibson et al., 2009). Kolibab et al. (2005a) also suggest that changes in Ig gene usage may account for the different affinity of vaccine-specific antibodies between young and older populations. Thus there may be a causal relationship between the distortion of B cell repertoires and vaccine hypo-responsiveness with age. However, previous repertoire analyses to investigate similar subjects are restricted to certain genes and isotypes due to the small numbers of sequences available (Kolibab et al., 2005a; Smithson et al., 2005).

We have performed high-throughput sequence analysis of *IGH* repertoires in human peripheral blood in order to investigate B cell responses following vaccination with the pneumococcal and influenza vaccines. In contrast to the oligoclonal response previously reported in response to the pneumococcal vaccine (Zhou et al., 2004; Kolibab et al., 2005a), here we report a very diverse and resilient vaccine response, with increased clone sizes at D7 and normal repertoire resumed at D28 (Figures 2 and 5). The discrepancy in diversity with previous reports may reflect the number of sequences analyzed and/or the fact that the influenza vaccine was included in this study in addition to the pneumococcal vaccine. Since the peak of the plasmablast response is generally between days 6 and 8 (Cox et al., 1994; Wrammert et al., 2008), and plasma cells and plasmablasts have many more copies of

immunoglobulin RNA per cell than B cells (Kuo et al., 2007), it is reasonable to assume that the clonal expansions seen at D7 represent the cells that are responding to vaccine. We can also see vaccine-induced changes in the CDR-H3 repertoire despite the fact that there are a large number of different antigens in this challenge. We previously showed that memory B cells in general have shorter CDR-H3 regions that are more hydrophilic (Wu et al., 2010), implying that Ig genes with these characteristics are preferentially selected in many different responses. These changes are also seen at D7 here (Figures 1E,F) and these data together strongly imply that the antibody immune response to challenge is skewed toward antibodies with certain characteristics of the antigen binding region even though the antigens themselves can be quite variable.

Many studies use specific serum IgG levels as a correlate of vaccine protection, although this metric is not always the best indicator. There are no age-related differences in anti-PPS IgG serology, but there is a significant age-related decrease in serum antibody function as determined by the opsonophagocytic assay (Anttila et al., 1999). Recent evidence suggests that this is due to a decrease in IgM antibody, since depletion of IgM from serum results in decreased opsonophagocytic activity (Park and Nahm, 2011; Sasaki et al., 2011). Our repertoire analysis does not show any age-related changes in IgG, except for a higher mutational frequency in the older group that is likely due to longer prior exposures to challenge with age. Nor do we see any significant age-related differences in the level of clonal expansion of IgM sequences. We do, however, see significant age-related differences

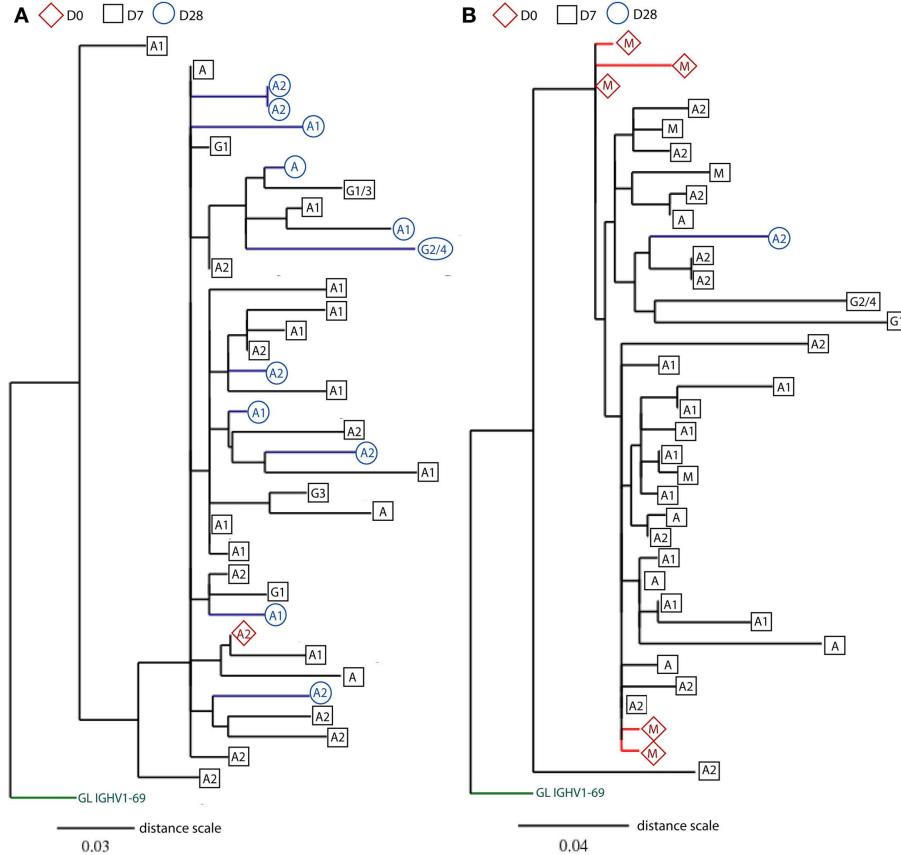


FIGURE 6 | Rooted phylogenograms show clonally related *IGH* sequences sampled at different days. CDR-H3 sequence motifs were used to identify clone relatives of *IGH* sequences found across D0, D7, and D28. After editing and aligning *IGH* sequences with germline gene using DNASTAR software, mutational phylogenetic trees were generated using the online Phylogeny Analysis program (Dereeper et al., 2010). Leaves represent *IGH* sequences sampled at different days (D0: red diamonds, D7: black squares, and D28: blue circles), and the putative germline (GL) genes are indicated at the base of

trees in green. The class (IgG, IgA, and IgM) or subclass (indicated by numbers) of *IGH* genes is indicated in the shapes. Internal nodes represent hypothetical intermediates of *IGH* sequences, and the branch lengths indicate percentage mutation differences between *IGH* sequences in proportion to the distance scale. Only two illustrations of phylogenetic trees are shown (**A**) from an old donor aged 71 years and (**B**) from a young donor aged 22 years. For space consideration, sequences in the two clones are randomly chosen for illustration.

in the CDR-H3 characteristics (Figure 4) and levels of mutation in the IgM repertoire (Figure 3), with generally less mutation and longer CDR-H3 in the old (both these factors being more characteristic of naïve B cell repertoires rather than memory B cell repertoires). The differences were significant at all time points so may not necessarily be confined to a change in the response. The IgM repertoire includes the naïve B cell population as well as IgM memory cells, so a change in the proportions of these two populations would also have an effect on our repertoire observations; thus, a decrease in IgM memory cells such as has been previously suggested may account for some difference (Shi et al., 2005). However, both age groups did show an increased level of mutation at D7 and a decrease at D28, although to a lesser extent in the old group. In order to look at the differences between age groups in the response without the background of naïve cells we split the data to look at sequences that were part of large clones in isolation, on the assumption that if a clone has expanded it is part of the IgM memory response rather than the naïve B cells in the background. The older IgM response repertoire had longer CDR-H3

and less mutation (Figures 3 and 4), which would strongly suggest that there is an age-related defect in the normal mechanisms of selection and hypermutation in IgM memory. We do not know what the specificity of these expanded IgM clones is, although based on previous literature we would hypothesize that they are T-independent responders to the polysaccharide antigen (Lortan et al., 1992). IgG2 antibodies are also thought to respond to T-independent polysaccharide antigens (Lortan et al., 1992) so it is interesting that alongside the defect in IgM memory repertoire there is a skewing in favor of IgG2 use in the IgG repertoire of older people.

A role for IgA in the protective vaccine response against influenza and pneumonia has not previously been highlighted, and the removal of IgA from serum was not shown to have any effect on the opsonophagocytic capability of post-vaccine serum. Since these respiratory diseases originate at mucosal surfaces it would seem plausible that IgA has some vital function, even if not in the circulation. It is clear from our data that there are significant differences in the IgA response in older people. In a

similar manner to the IgM memory cells there is less hypermutation in the expanded clones and they have larger CDR-H3 regions. The degree of clonal expansion is less overall and takes much longer, to the extent that it still appears to be occurring at D28 after the younger group has contracted the response. The IgA serum antibody response is short lived. In contrast to IgG and IgM antibodies, which increase in serum with maximum values at D28, the maximal serum level of IgA is at D7 in the normal young population and it decreases by D28 (Ademokun et al., 2011). The clonal expansions of all three isotypes peaks at D7 and contracts at D28 in the blood (Figure 2B). Thus for IgG and IgM one can envisage a scenario where cells have left the blood to reside in a niche where they continue to secrete antibody into the circulation. However the serum IgA antibody concentration seems to mirror the clonal expansion data, which may indicate that the cells are short lived and do not go on to secrete antibody in survival niches, or perhaps they do survive but secrete IgA antibody at mucosal surfaces rather than into the circulation.

Since the most striking age-related difference in these data was in the IgA response we split the data into IgA1 and IgA2 sequences. IgA1 has previously been associated with serum responses and IgA2 with mucosal responses (Russell et al., 1992). It is clear from our lineage tree analysis that IgA1 and IgA2 can be quite closely related, since we find many clones containing both subclasses (Figure 6). However in Figure 5C we see that the main age-related difference in clonal expansions that we saw at D7 and D28 is mainly due to IgA1 rather than IgA2, and we also find clones which do not mix the subclasses, so it is possible that certain types of antigens/responses may elicit one subclass only.

The existence of many clones with relatives in both pre- and post-vaccination samples (Table 2) indicates that some of these clones may be very large even before vaccination. The chances of finding a particular clonotype in a sample are dependent on the number of cells sampled, the total number of cells in the blood, and the total number of clone members in the blood. We would not expect to find a particular clonotype more than once if it were not part of an expanded clone. In the simplest terms, if we assume that there are 10^8 B cells in the blood, and we have found two related sequences in a sample of 5000, then there could be approximately $(10^8/5000)^2$, or 40,000, related sequences in that one clone in the blood altogether. Also, if we assume that an expansion at D7 originated from a single unique precursor in the blood at D0, the chances of sampling and sequencing that precursor would be 1 in 10^8 . Since we see nearly 400 examples of clones where there are related sequences at D0 then we can provisionally conclude that many expanded precursors of cells with specificity for these vaccines are already present in the blood pre-challenge. Whether

these pre-existing specificities are mono-specific for the antigens in the vaccine or are cross-reactive from a prior, related, challenge cannot be determined since it was impossible to determine the prior extent of exposure to influenza or *S. pneumoniae* in the participants. These high-throughput data have also shown that we need to be careful about interpretation of lineage trees with respect to inference of chronological ordering since the trees can be quite complex (Figure 6). If there has been extensive prior expansion of cells in the blood it cannot be assumed that all clones in the expansion will have mutated at the same rate or in the same reaction. Hence a random sampling will not always result in the samples from the earlier time points appearing at the top of the lineage trees (i.e., with less mutations from germline). Similarly, extensive expansion followed by random switching and sampling may result in a seemingly impossible succession of switching events, such as *IGHM* sequences appearing downstream of *IGHA* sequences. However this could simply mean that a cell has expanded without mutation, 50% of them have switched and the sampling has picked up one of the switched parents together with offspring of one of the unswitched parents.

In conclusion, our high-throughput *IGH* repertoire analyses have demonstrated that we can visualize an immune response to vaccine by the expansion of clonotypes expressing particular Ig genes, and with particular CDR-H3 characteristics. The large clonal expansions indicate a complex recall response. There are significant age-related differences in the response with respect to subclass distribution, particularly in the extent and timing of IgA clonal expansion and skewing toward greater use of IgG2. Older responding IgM and IgA clonotypes are also less mutated and use a longer CDR-H3, which might affect antigen recognition. Although much more now needs to be done to explore the significance of the age-related changes in IgA responses described here, our work does highlight the critical need to consider different classes and subclasses of antibody in vaccine studies in general.

AUTHOR CONTRIBUTION

Yu-Chang Bryan Wu designed and carried out experiments, analyzed data and wrote the manuscript; David Kipling designed and ran the data handling and analysis scripts, analyzed data, and wrote the manuscript; Deborah K. Dunn-Walters, oversaw the project, designed experiments and analytical tools, carried out data analysis and wrote the manuscript.

ACKNOWLEDGMENTS

This work was funded by the Human Frontiers Science program and Research into Aging, Age UK. The authors would like to thank all our volunteers for their blood donation.

REFERENCES

- Ademokun, A., Wu, Y. C., Martin, V., Mitra, R., Sack, U., Baxendale, H., Kipling, D., and Dunn-Walters, D. K. (2011). Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* 10, 922–930.
- Anttila, M., Voutilainen, M., Jantti, V., Eskola, J., and Kayhty, H. (1999).
- Banerjee, M., Mehr, R., Belelovsky, A., Spencer, J., and Dunn-Walters, D. K. (2002). Age- and tissue-specific differences in human germinal center B cell selection revealed by analysis of IgVH gene hypermutation and lineage trees. *Eur. J. Immunol.* 32, 1947–1957.
- Arzt, A. S., Ersler, W. B., and Longo, D. L. (2003). Pneumococcal vaccination and revaccination of older adults. *Clin. Microbiol. Rev.* 16, 308–318.
- Baxendale, H. E., Keating, S. M., Johnson, M., Southern, J., Miller, E., and Goldblatt, D. (2010). The majority of intestinal IgA+ and IgG+ plasmablasts in the early kinetics of circulating pneumococcal-specific memory B cells following pneumococcal conjugate and plain polysaccharide vaccines in the elderly. *Vaccine* 28, 4763–4770.
- Benckert, J., Schmolka, N., Kreschel, C., Zoller, M. J., Sturm, A., Wiedemann, B., and Wardemann, H. (2011). The majority of intestinal IgA+ and IgG+ plasmablasts in the

- human gut are antigen-specific. *J. Clin. Invest.* 121, 1946–1955.
- Brochet, X., Lefranc, M. P., and Giudicelli, V. (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 36, W503–W508.
- Brown, T. A., Murphy, B. R., Radl, J., Haaijman, J. J., and Mestecky, J. (1985). Subclass distribution and molecular form of immunoglobulin A hemagglutinin antibodies in sera and nasal secretions after experimental secondary infection with influenza A virus in humans. *J. Clin. Microbiol.* 22, 259–264.
- Carson, P. J., Schut, R. L., Simpson, M. L., O'Brien, J., and Janoff, E. N. (1995). Antibody class and subclass responses to pneumococcal polysaccharides following immunization of human immunodeficiency virus-infected patients. *J. Infect. Dis.* 172, 340–345.
- Cox, R. J., Brokstad, K. A., Zuckerman, M. A., Wood, J. M., Haaheim, L. R., and Oxford, J. S. (1994). An early humoral immune response in peripheral blood following parenteral inactivated influenza vaccination. *Vaccine* 12, 993–999.
- Dereeper, A., Audic, S., Claverie, J. M., and Blanc, G. (2010). BLAST-EXPLORER helps you building datasets for phylogenetic analysis. *BMC Evol. Biol.* 10, 8. doi:10.1186/1471-2148-10-8
- Frasca, D., Landin, A. M., Lechner, S. C., Ryan, J. G., Schwartz, R., Riley, R. L., and Blomberg, B. B. (2008). Aging down-regulates the transcription factor E2A, activation-induced cytidine deaminase, and Ig class switch in human B cells. *J. Immunol.* 180, 5283–5290.
- Gibson, K. L., Wu, Y. C., Barnett, Y., Duggan, O., Vaughan, R., Kondeatis, E., Nilsson, B. O., Wikby, A., Kipling, D., and Dunn-Walters, D. K. (2009). B-cell diversity decreases in old age and is correlated with poor health status. *Aging Cell* 8, 18–25.
- Hannoun, C., Megas, F., and Piercy, J. (2004). Immunogenicity and protective efficacy of influenza vaccination. *Virus Res.* 103, 133–138.
- Hocart, M. J., Mackenzie, J. S., and Stewart, G. A. (1990). Serum IgG subclass responses of humans to inactivated and live influenza A vaccines compared to natural infections with influenza A. *J. Med. Virol.* 30, 92–96.
- Ikai, A. (1980). Thermostability and aliphatic index of globular proteins. *J. Biochem.* 88, 1895–1898.
- Kolibab, K., Smithson, S. L., Rabquer, B., Khuder, S., and Westerink, M. A. (2005a). Immune response to pneumococcal polysaccharides 4 and 14 in elderly and young adults: analysis of the variable heavy chain repertoire. *Infect. Immun.* 73, 7465–7476.
- Kolibab, K., Smithson, S. L., Shriner, A. K., Khuder, S., Romero-Steiner, S., Carbone, G. M., and Westerink, M. A. (2005b). Immune response to pneumococcal polysaccharides 4 and 14 in elderly and young adults. I. Antibody concentrations, avidity and functional activity. *Immun. Ageing* 2, 10.
- Kuo, T. C., Shaffer, A. L., Haddad, J. Jr., Choi, Y. S., Staudt, L. M., and Calame, K. (2007). Repression of BCL-6 is required for the formation of human memory B cells in vitro. *J. Exp. Med.* 204, 819–830.
- Lindner, C., Wahl, B., Fohse, L., Suerbaum, S., Macpherson, A. J., Prinz, I., and Pabst, O. (2012). Age, microbiota, and T cells shape diverse individual IgA repertoires in the intestine. *J. Exp. Med.* 209, 365–377.
- Lortan, J. E., Vellodi, A., Jurges, E. S., and Hugh-Jones, K. (1992). Class- and subclass-specific pneumococcal antibody levels and response to immunization after bone marrow transplantation. *Clin. Exp. Immunol.* 88, 512–519.
- Lue, C., Tarkowski, A., and Mestecky, J. (1988). Systemic immunization with pneumococcal polysaccharide vaccine induces a predominant IgA2 response of peripheral blood lymphocytes and increases of both serum and secretory anti-pneumococcal antibodies. *J. Immunol.* 140, 3793–3800.
- McCullers, J. A. (2006). Insights into the interaction between influenza virus and pneumococcus. *Clin. Microbiol. Rev.* 19, 571–582.
- Nichol, K. L., Nordin, J. D., Nelson, D. B., Mullooly, J. P., and Hak, E. (2007). Effectiveness of influenza vaccine in the community-dwelling elderly. *N. Engl. J. Med.* 357, 1373–1381.
- Osterholm, M. T., Kelley, N. S., Sommer, A., and Belongia, E. A. (2012). Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. *Lancet Infect. Dis.* 12, 36–44.
- Park, S., and Nahm, M. H. (2011). Older adults have a low capacity to opsonize pneumococci due to low IgM antibody response to pneumococcal vaccinations. *Infect. Immun.* 79, 314–320.
- Powers, D. C. (1994). Effect of age on serum immunoglobulin G subclass antibody responses to inactivated influenza virus vaccine. *J. Med. Virol.* 43, 57–61.
- Romero-Steiner, S., Musher, D. M., Cetron, M. S., Pais, L. B., Groover, J. E., Fiore, A. E., Plikaytis, B. D., and Carbone, G. M. (1999). Reduction in functional antibody activity against *Streptococcus pneumoniae* in vaccinated elderly individuals highly correlates with decreased IgG antibody avidity. *Clin. Infect. Dis.* 29, 281–288.
- Russell, M. W., Lue, C., Van Den Wall Bake, A. W., Moldoveanu, Z., and Mestecky, J. (1992). Molecular heterogeneity of human IgA antibodies during an immune response. *Clin. Exp. Immunol.* 87, 1–6.
- Sanal, O., Ersoy, F., Yel, L., Tezcan, I., Metin, A., Ozayrek, H., Gariboglu, S., Fikrig, S., Berkel, A. I., Rijkers, G. T., and Zegers, B. J. (1999). Impaired IgG antibody production to pneumococcal polysaccharides in patients with ataxia-telangiectasia. *J. Clin. Immunol.* 19, 326–334.
- Sasaki, S., Sullivan, M., Narvaez, C. F., Holmes, T. H., Furman, D., Zheng, N. Y., Nishtala, M., Wrammert, J., Smith, K., James, J. A., Dekker, C. L., Davis, M. M., Wilson, P. C., Greenberg, H. B., and He, X. S. (2011). Limited efficacy of inactivated influenza vaccine in elderly individuals is associated with decreased production of vaccine-specific antibodies. *J. Clin. Invest.* 121, 3109–3119.
- Shi, Y., Yamazaki, T., Okubo, Y., Uehara, Y., Sugane, K., and Agematsu, K. (2005). Regulation of aged humoral immune defense against pneumococcal bacteria by IgM memory B cell. *J. Immunol.* 175, 3262–3267.
- Simell, B., Kilpi, T., and Kayhty, H. (2006). Subclass distribution of natural salivary IgA antibodies against pneumococcal capsular polysaccharide of type 14 and pneumococcal surface adhesin A (PsaA) in children. *Clin. Exp. Immunol.* 143, 543–549.
- Smithson, S. L., Kolibab, K., Shriner, A. K., Srivastava, N., Khuder, S., and Westerink, M. A. (2005). Immune response to pneumococcal polysaccharides 4 and 14 in elderly and young adults: analysis of the variable light chain repertoire. *Infect. Immun.* 73, 7477–7484.
- Szabo, P., Li, F., Mathew, J., Lillvis, J., and Weksler, M. E. (2004). Evolution of B-cell clonal expansions with age. *Cell. Immunol.* 231, 158–167.
- Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C., Williams, K. L., Appel, R. D., and Hochstrasser, D. F. (1999). Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* 112, 531–552.
- Wrammert, J., Smith, K., Miller, J., Langley, W. A., Kokko, K., Larsen, C., Zheng, N. Y., Mays, I., Garman, L., Helms, C., James, J., Air, G. M., Capra, J. D., Ahmed, R., and Wilson, P. C. (2008). Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* 453, 667–671.
- Wu, Y. C., Kipling, D., Leong, H. S., Martin, V., Ademokun, A. A., and Dunn-Walters, D. K. (2010). High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* 116, 1070–1078.
- Wu, Y.-C. B., Kipling, D., and Dunn-Walters, D. K. (2011). The relationship between CD27 negative and positive B cell populations in human peripheral blood. *Front. Immunol.* 2:81. doi:10.3389/fimmu.2011.00081
- Zhou, J., Lottenbach, K. R., Barenkamp, S. J., and Reason, D. C. (2004). Somatic hypermutation and diverse immunoglobulin gene usage in the human antibody response to the capsular polysaccharide of *Streptococcus pneumoniae* type 6B. *Infect. Immun.* 72, 3505–3514.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 March 2012; accepted: 20 June 2012; published online: 09 July 2012.

Citation: Wu Y-CB, Kipling D and Dunn-Walters DK (2012) Age-related changes in human peripheral blood IGH repertoire following vaccination. *Front. Immunol.* 3:193. doi:10.3389/fimmu.2012.00193

This article was submitted to Frontiers in B Cell Biology, a specialty of Frontiers in Immunology.

Copyright © 2012 Wu, Kipling and Dunn-Walters. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



Natural and man-made V-gene repertoires for antibody discovery

William J. J. Finlay¹ and Juan C. Almagro^{2*}

¹ Global Biotherapeutics Technologies, Pfizer, Dublin, Ireland

² Centers for Therapeutic Innovation, Pfizer, Boston, MA, USA

Edited by:

Harry W. Schroeder, University of Alabama at Birmingham, USA

Reviewed by:

Laurence Morel, University of Florida, USA

Wenxia Song, University of Maryland, USA

***Correspondence:**

Juan C. Almagro, CTI-Boston, Pfizer,
3 Blackfan Circle - 18th Floor,
Boston, MA 02115, USA.

e-mail: juan.c.almagro@pfizer.com

Antibodies are the fastest-growing segment of the biologics market. The success of antibody-based drugs resides in their exquisite specificity, high potency, stability, solubility, safety, and relatively inexpensive manufacturing process in comparison with other biologics. We outline here the structural studies and fundamental principles that define how antibodies interact with diverse targets. We also describe the antibody repertoires and affinity maturation mechanisms of humans, mice, and chickens, plus the use of novel single-domain antibodies in camelids and sharks. These species all utilize diverse evolutionary solutions to generate specific and high affinity antibodies and illustrate the plasticity of natural antibody repertoires. In addition, we discuss the multiple variations of man-made antibody repertoires designed and validated in the last two decades, which have served as tools to explore how the size, diversity, and composition of a repertoire impact the antibody discovery process.

Keywords: therapeutic antibodies, antigen-binding site, antibody structure, structure-function relationship

INTRODUCTION

In recent decades, rodent monoclonal antibodies obtained by hybridoma technology and engineered by molecular biology techniques, or human antibodies obtained by display technologies or B-cell cloning, have become the treatment of choice in diverse diseases such as multiple sclerosis, rheumatoid arthritis, and several types of cancers, making a significant component of the pharmaceuticals market (Nelson et al., 2010). The success of therapeutic antibodies, with as many as 28 antibodies and antibody fragments marketed in The United States or The European Union (Reichert, 2012), resides in their exquisite specificity, high potency, stability, solubility, clinical tolerability, and relatively inexpensive manufacturing process in comparison with other biologics.

The factors contributing to the specificity and potency of antibodies have intrigued scientists since their discovery in the late 1800s and only in the last three decades has a clear picture of how antibodies work emerged. The current knowledge base has been assembled by combining insights from multiple disciplines such as: structural biology—studying hundreds of x-ray crystallography antibody structures from different species (Davies and Metzger, 1983; Chothia and Lesk, 1987; Wilson and Stanfield, 1994; Stanfield and Wilson, 2010) free and in complex with a wide variety of ligands (MacCallum et al., 1996; Ragunathan et al., 2012); immunogenetics—by fully characterizing the germline gene antibody repertoire of humans and other species (Lefranc et al., 2005) and by deciphering the molecular mechanisms used to generate functional antibody molecules starting from diverse gene families (Tonegawa, 1983); and cellular immunology—dissecting the process by which *in vivo* selection of specific antibodies occurs during an immune response and understanding the mechanisms that allow the affinity and

specificity of the selected antibodies to mature as the immune response progresses (Noia and Neuberger, 2007).

The accumulation of this knowledge has potentiated several technological advances in the antibody engineering field, such as humanization of non-human antibodies to increase their human content and to enhance their manufacturability profile (Gilliland et al., 2012), the development of display technologies to select specific human antibodies *in vitro* (Hoogenboom, 2005), and the engineering of antibody characteristics such as affinity, cross-reactivity with target orthologs, stability, and solubility. Each of these great leaps forward have relied directly on a core of fundamental immunological knowledge and made it possible to create close to 30 antibody-based drugs, at the time of writing.

Here, we first provide an overview of the antibody structure and outline the fundamental principles that define how antibodies interact with diverse ligands. In the second section, we review the current knowledge of the antibody repertoire of humans and experimental species commonly used to generate monoclonal antibodies such as mice, chickens, and camelids. Each of these species possess distinct germline gene repertoires, have differing mechanisms of generating and affinity maturing their antibody molecules and, therefore, offer alternative sources of specific variable regions for therapeutic antibody development. In the third section, multiple variations of man-made antibody repertoires are described, from their inception to the current state of the art. These designer repertoires have applied the compound knowledge derived from both structural and repertoire studies, serving as tools to test hypotheses on how the size of a repertoire, its diversity and composition impact the selection of more specific and higher affinity antibodies. These repertoires have also been used extensively by academic laboratories and biotech companies to discover and optimize human antibodies *in vitro*. At the end

of the article, a section with conclusions and future directions is included.

THE ANTIBODY MOLECULE

The IgG isotype is the most abundant form of circulating antibody and the molecular format of choice for most marketed therapeutic antibodies (Reichert, 2012), as it is stable, soluble, readily expressed in heterologous systems such as Chinese hamster ovary (CHO) cells and can potentially engage effector functions such as antibody-dependent cell-mediated cytotoxicity (ADCC) and complement-dependent cytotoxicity (CDC). IgGs are Y-shaped glycoproteins of approximately 150 kDa composed of two identical polypeptide heavy (H) chains and two identical light (L) chains. The most abundant classes of L chains are κ and λ , which are functionally indistinguishable, but structurally different and vary in proportion in different species. For instance, the human repertoire is approximately 40:60 λ : κ , whereas, the mouse repertoire is \sim 95% κ -type. The H chain divides IgGs into five classes, IgG, IgD, IgE, IgA, and IgM, each with a unique role in the adaptive immune system.

By digesting IgGs with papain, two fractions can be obtained, one containing the so-called crystallizable fragment (Fc) and the other containing two identical antigen-binding fragments or Fabs (Figure 1). In the Fc resides the effector functions, whereas, the Fab, as its name indicates, binds the antigen and thus defines the specificity of antibodies. Each Fab has two variable domains, one from the H chain (V_H) and another from the L chain (V_L), in addition to two C domains: C_{H1} and C_{L1} . The Fc is a dimer made of four C domains, two C_{H2} and two C_{H3} domains.

The first antibody structures, solved in the 1970s [for early reviews see (Padlan, 1977; Amzel and Poljak, 1979; Davies and Metzger, 1983) and for more current reviews see (Wilson and Stanfield, 1994; Stanfield and Wilson, 2010)], revealed that

V- and C-domains have a conserved and similar structure, termed “immunoglobulin (Ig) fold.” The Ig fold is also the building block of a large number of other proteins with diverse functions, which are collectively called the Ig superfamily (Williams and Barclay, 1988). The Ig fold consists of two anti-parallel β -sheets that are tightly packed together. In the C domain, one of the β -sheets is formed by four β -strands A to D, whereas, the other β -sheet is formed by three β -strands C to G (Figure 2). A conserved intra-domain disulfide bridge, formed between cysteine residues in the B and F β -strands, stabilizes the C domain. The V-domain have an insertion with respect to the C domain of two extra β -strands, identified as C' and C'', present between β -strands C and D (Figure 2). As in the C domain, an intra-domain disulfide bridge is formed between cysteine residues in β -strands B and F. The V-domains are in general less compact than the C domains with some longer loops connecting the β -strands. This flexibility and the longer loops contribute to the mechanism of antigen binding, thus defining the capability of antibodies to recognize diverse antigens.

THE ANTIGEN-BINDING SITE

The antigen binding site is principally defined by the Complementarity-Determining Regions (CDRs). These regions were originally identified by amino acid sequence variability analysis (Wu and Kabat, 1970; Kabat and Wu, 1971) as highly variable regions within the V-domains. The CDRs were defined prior to our knowledge of the mechanisms by which antibodies are generated and predated the three-dimensional structure solution of antibodies. Once the first Fab structures were solved, it was realized that the CDRs approximately correspond to loops that vary in structure, called hypervariable loops (HVLs). Each V-domain contributes three CDRs to the antigen-binding site: CDR-L1, CDR-L2, and CDR-L3 from the V_L and CDR-H1,

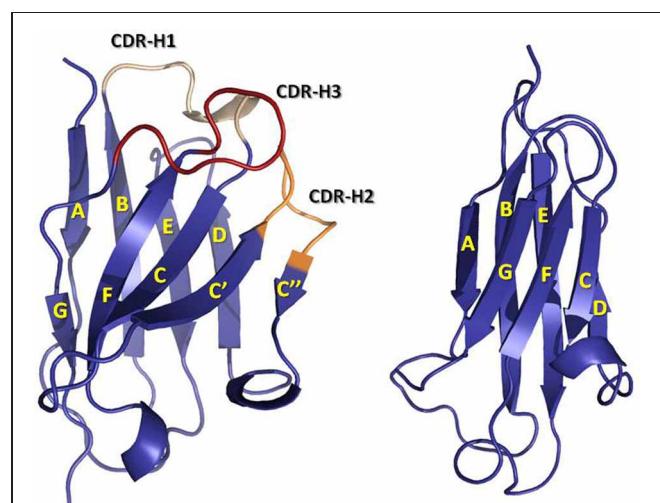
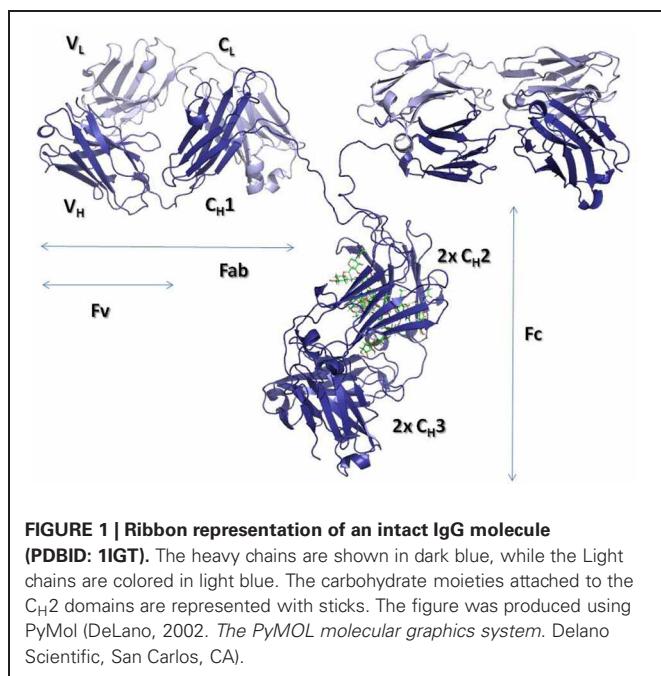


FIGURE 2 | Ribbon representation of a V_H (left) domain and a C_{H1} (right) domain. CDRs are colored in yellow (CDR-1), orange (CDR-2), and red (CDR-3). Note the insertion in the V_H domain with respect to the C_{H1} domain of two β -strands, C' and C'', and the loop linking them, which contains the CDR-H2. The coordinates used to produce the Figure were the same as in Figure 1. The figure was generated with PyMol.

CDR-H2, and CDR-H3 from the V_H . The three CDRs from V_H and the three from V_L are brought together by non-covalent association of the V-domains at the N-terminal region of the Fv (**Figure 3**). The remaining portion of the V-domain, i.e., the two β -sheets and non-HVLs, generally provide structural support to the antigen-binding site, rather than making contact with antigen and are thus referred to as framework regions (FRs). However, the sequence variability observed in the FRs is not irrelevant to functional binding diversity, as it can directly affect CDR loop conformation and the orientation of V_H - V_L pairing (Foote and Winter, 1992; Abhinandan and Martin, 2010).

Given the essential variability of the antigen-binding site, which must be capable of recognizing a large array of diverse antigens to fulfill its remit, it was initially thought that each antibody possesses a unique conformation at the antigen-binding site. Nevertheless, analysis (Chothia and Lesk, 1987; Chothia et al., 1989) in the late 1980s of a small set of structures of immunoglobulin fragments available at the time revealed that, although the HVLs vary in sequence, five out of the six HVLs (CDR-L1, CDR-L2, CDR-L3, CDR-H1, and CDR-H2) had a limited set of main-chain conformations or “canonical structures.” The canonical structure model implied a paradigm shift in the field, replacing the notion that each antibody has unique HVL conformations and thus overall unique antigen-binding site structure. The limited set of canonical structures helped to develop 3D modeling structure strategies (Martin and Thornton, 1996) and suggested that structural constraints are at work in antigen recognition.

A canonical structure is defined by the HVL length and conserved residues located in the HVL and FR (Chothia and Lesk, 1987). Overall, the structural repertoire generated by λ -type chains is broader than that of κ -type chains (Chailyan et al., 2011). In the latter, all the canonical structures at CDR-L1 follow a similar pattern, which consists of an extended conformation between residues 26 and 29 [Chothia’s numbering;

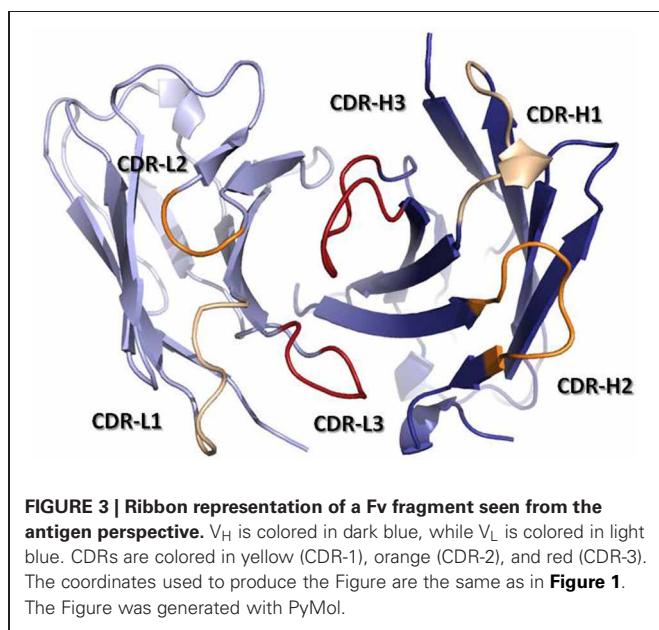
(Al-Lazikani et al., 1997)], and hairpin loops of different lengths encompassing residues 30–32, with up to seven insertions in this segment of the loop. The CDR-L2 adopts a single conformation. Most (~70%) of the CDR-L3 loops have a single canonical structure. In contrast, the CDR-L1 of λ -type chains adopts a helical structure with up to eight conformations and an average root mean square deviation (RMSD) between loops of 2.3 Å (Chailyan et al., 2011). The λ -type CDR-L2 usually adopts a similar hairpin loop conformation to that of κ -type, but can in some instances have an insertion of four residues, which leads to another canonical structure. The CDR-L3 in λ -type antibodies has a broader variety of lengths and conformations than κ -type antibodies, with only a small fraction of the loops following a defined canonical structure (Chailyan et al., 2011).

The CDR-H1, similar to the CDR-L1, has an extended conformation linking β -strands from the two β -sheets that form the Ig fold. However, it is less diverse than its counterpart in V_L , with three canonical structures and a strong bias (~85%) (Ragunathan et al., 2012) toward the shortest loop (seven residues). The repertoire of canonical structures of CDR-H2 is less skewed than that for CDR-L3 and CDR-H1, with six canonical structures. Still, 59–70% of the antibodies (Ragunathan et al., 2012) have a six-residue canonical structure.

Recently, the application of clustering algorithms (North et al., 2011) on 300 non-redundant antibody structures has further stratified the canonical structure combinations by identifying 28 HVL combinations of lengths for the loops with canonical structures, whereas, previous analysis (Al-Lazikani et al., 1997) covered only 20. Only four of these clusters had more than one conformation, of which two could be distinguished by gene source (mouse/human; κ/λ) and one could be distinguished solely by the presence and position of Proresidues in the CDR-L3. Of the 28 CDR-lengths, 15 have multiple conformational clusters, including 10 for which previous analysis had only one canonical structure combination.

The CDR-H3, localized at the center of the antigen-binding site, is by far the most variable loop in length and sequence of the CDRs (Chothia and Lesk, 1987; Wu et al., 1993; Zemlin et al., 2003). The diversity of the CDR-H3 comes from the recombination of three germline genes: IGHV,IGHD, and IGHJ (Tonegawa, 1983), imprecise recombination of these genes, i.e., junctional diversity (Alt and Baltimore, 1982), the possibility of using three reading frames for translation of the IGHD gene (Sanz, 1991), and further diversification during somatic hypermutation process (see below).

Human CDR-H3 loops have an average length of 15.2 (± 4.1) residues (IMGT CDR definition) (Zemlin et al., 2003), with a range of lengths between 1–35 residues, and a length distribution resembling a Gaussian process. While extensive analysis of antibody structures has identified sequence patterns to predict the conformation of the residues at the base of the CDR-H3 (Shirai et al., 1996; Morea et al., 1998), the enormous variability in amino acid sequence and length of this loop, as well as its flexibility, has precluded delineation of rules for predicting its overall conformation. Thus, structural modeling of CDR-H3 is still challenging (Almagro et al., 2011), using either comparative



methods that rely on templates chosen based on sequence homology, or knowledge-based methods such as the canonical structure model.

STRUCTURE-FUNCTION RELATIONSHIPS AT THE ANTIGEN-BINDING SITE

Since antibodies have a small subset of canonical structures in five of the six loops that define the antigen-binding site, it is reasonable to hypothesize that only a limited subset of antigen-binding site geometries exists, and the arising questions are whether the general architecture of the antigen-binding site can be predicted and whether it correlates with antigen recognition (Vargas-Madrazo et al., 1995). Finding structure-function correlations at the antigen-biding site holds the promise of providing insights into the mechanism of the molecular recognition process used by antibodies to bind diverse antigens and thereby to assist the rational design of antibodies of desired specificity.

Initial work (Vargas-Madrazo et al., 1995) showed that from a total of 300 possible canonical structure combinations described at that time, only 10 exist in 90% of the sequences analyzed. The existing canonical structure combinations were classified in two sets: one with preference for some specific types of antigens like proteins, peptides or haptens, and other with multi-specific binding capabilities. In the specific classes, the length of CDR-H2 and CDR-L1 was found to correlate with the type of antigen, whereas, in the multi-specific classes, such a correlation could not be established. A recent study (Ragunathan et al., 2012) of 140 unique antigen-antibody complexes has corroborated that most of the anti-protein antibodies have canonical structures determined by short CDR-L1 loops (6–8 residues). This is in contrast to anti-peptide and anti-hapten antibodies, which predominantly have canonical structures made of long CDR-L1 loops (11–13 residues). The remaining loops show little difference in the canonical structure distribution across anti-protein, anti-peptide, and anti-hapten antibodies.

Figure 4 overlays 99 unique mid to high resolution ($\leq 3.0 \text{ \AA}$) antibody structures, including 30 in complex with proteins, 34 with peptides, and 35 with haptens. As can be seen, the topography of the antigen-binding site tends to determine the size of the antigen with which the antibody interacts. Anti-protein

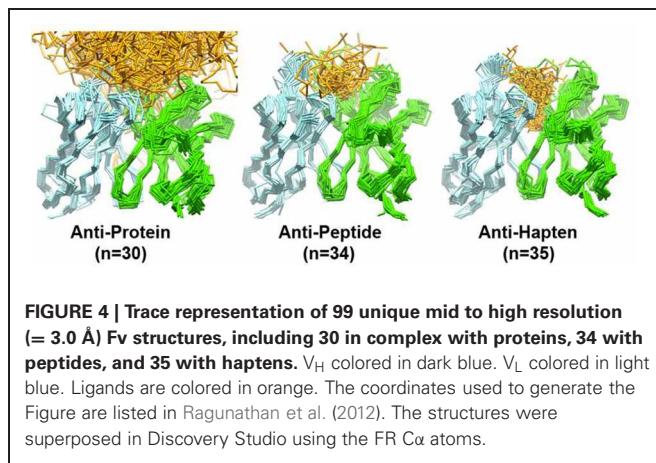
antibodies tend to have flatter binding sites than anti-peptide antibodies. The antigen-binding of anti-peptide antibodies is grooved, mainly determined by the long CDR-L1, which accommodates the peptides at the center of the antigen-binding site. Anti-hapten antibodies have a smaller antigen-biding site with contacts with haptens being buried deeper in the $V_H:V_L$ interface where proteins and peptides cannot reach.

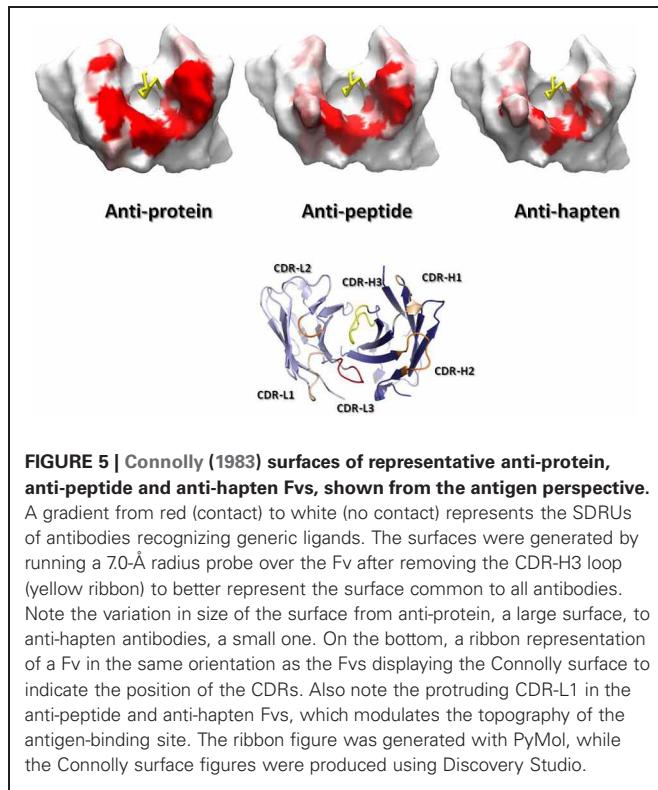
In non-specific classes, the CDR-H3 plays a predominant role in defining the topography of the binding site (Vargas-Madrazo et al., 1995). Short CDR-H3 loops can create a cavity in the antigen-binding site to accommodate peptides. Long CDR-H3 loops are found in antibodies associated with chronic viral infections, in contrast to antibodies from acute viral infections, which have relatively short CDR-H3 loops (Breden et al., 2011). Long and extended CDR-H3 loops can generate a definite type of structure, called finger-like topography (Saphire et al., 2001), which differ from the typical flat anti-protein binding-site. This finger-like topography allows antibodies to access recessed epitopes in the viral proteins.

The number of residues in contact with antigens also differs in antibodies recognizing proteins, peptides and haptens (MacCallum et al., 1996; Almagro, 2004; Ragunathan et al., 2012). The average number of contact residues in V_L for the anti-protein, anti-peptide, and anti-hapten antibodies is 9, 9, and 7, respectively (Ragunathan et al., 2012). The corresponding values for V_H are 14, 12, and 10. A more detailed analysis of the solvent accessible surface (SAS) that is buried upon antigen binding and the location and frequency of contacts (called specificity-determining residues usage, SDRUs) of antibodies in complex with proteins, peptides or haptens also show distinctive patterns (Almagro, 2004). Anti-protein antibodies have an average ($\pm SD$) SAS value of $737 (\pm 272) \text{ \AA}^2$ with hotspots of SDRUs located at the edge of the antigen-binding site (Ragunathan et al., 2012). Anti-hapten antibodies have a roughly 2-fold smaller SAS value of $374 (\pm 117) \text{ \AA}^2$ with hotspots of SDRUs placed in the interior of the antigen-binding site or even buried in the $V_L:V_H$ interface. Anti-peptide antibodies have a SAS value of $544 (\pm 158) \text{ \AA}^2$, which is in between anti-protein and anti-hapten antibodies. The SDRU hotspots of anti-peptide antibodies are located in the interior of the antigen-binding site but not buried in the $V_L:V_H$ interface as with anti-hapten antibodies.

Combining the SDRU patterns with the distinctive shape of the antigen-binding site of antibodies recognizing different types of antigens lead to the conclusion that anti-protein antibodies tend to have flatter and larger binding sites than anti-peptide and anti-hapten antibodies. The antigen-binding of anti-peptides is grooved, whereas, anti-hapten antibodies have a smaller and deeper antigen-biding site, with SDRU hotspots buried in the $V_H:V_L$ interface (**Figure 5**).

Since SDRUs are a measure of the likelihood of establishing contacts with the antigen, they can provide a definition of the antigen-binding site in absence of the antigen-antibody complex structure. A definition of the antigen-binding site based on SDRUs could thus guide the selection of residues to transfer the specificity from a given antibody into a different scaffold, either to produce a molecule with enhanced biophysical profile such as increased stability (Ewert et al.,





2004) and/or to humanize a nonhuman antibody (Almagro and Fransson, 2008). Importantly, protocols using SDRUs can tailor humanization of antibodies recognizing different types of ligands, thereby minimizing the region of the non-human antibody grafted into the human context and hence potential immunogenicity.

Not all amino acids are equally used to contact antigens and the types of antibody residues involved in contacts with proteins, peptides, and haptens also differ. Tyrosine (Y), arginine (R), asparagine (N), aspartic acid (D), histidine (H), serine (S), and threonine (T) make more contacts than other amino acids in all the three antigen types. Of particular interest is Y, which has been found in a high proportion in the antigen-binding site of antibodies. For instance, Lo Conte et al. (1999) observed that Y contributed to 16.6% of all amino acids in contact in the 19 antigen-antibody complexes available at that time. Similarly, an earlier report by Mian et al. (1991) based on the analysis of only six antibody–antigen complexes reported the overuse of Y to contact antigens. Cysteine (C), proline (P), glutamine (Q), glutamic acid (E) and hydrophobic amino acids such as alanine (A), valine (V), isoleucine (I), leucine (L), methionine (M) and phenylalanine (F) make significantly fewer contacts. Thus, hydrophilic amino acids predominate over hydrophobic ones. In the CDR-L1, N, and D are the most frequent residues in contacts. R is rare and tryptophan (W) does not occur. CDR-L2 has less diversity than CDR-L1 and CDR-L3, and in the latter, most contacts involve amino acids S, T, W, and Y, whereas, R, N, G, and H make fewer contacts.

A detailed analysis of the contribution of each amino acid called Specificity-Determining Residues Matrix (SDRM) to each SDRU depending upon the type of antigen the antibody interacts with have been described by Ragunathan et al. (2012). Briefly, there are more D and T contacts in anti-protein antibodies in CDR-L1 than anti-peptide and anti-hapten antibodies. In CDR-L3, anti-protein antibodies have more R and W contacts, whereas, anti-hapten antibodies have more Q, G, and H contacts. Similar to V_L, S, T, and Y dominate the contacts for V_H. Likewise, C, P, Q, D and hydrophobic amino acids are significantly underrepresented at contact residues. Interestingly, V_H has more contacts involving negatively charged amino acids and fewer K residues in comparison to V_L. In CDR-H1, N, G, S, T, and Y predominant in contact sites for all antibodies. In addition, D occurs frequently in anti-protein and anti-peptide antibodies. The detailed picture of the contribution of each amino acid (SDRM) to each SDRU depending upon the type of antigen the antibody interacts with has practical applications to design antibody repertoires.

THE ANTIBODY REPERTOIRE IN HUMANS AND IMMUNIZATION OF HOST SPECIES

In addition to the structural studies outlined above, antibody repertoire analyses and comparative immunogenetics have been highly informative approaches to understanding how antibodies evolved to recognize diverse antigen structures. Indeed, the lessons learned from such studies have been critical factors in the progress of antibody engineering. For example, a profound understanding of the biases inherent in the functional repertoire of human antibodies, in comparison to those of other species (Schroeder et al., 1995; Zemlin et al., 2003; Schroeder, 2006) inspired experimental work to define the critical biochemical characteristics required to form a functional synthetic antibody repertoire (Fellouse et al., 2006; Birtalan et al., 2008, 2010). To further illustrate the important influence these studies, below we outline what has been learned about the human antibody repertoire and several other species of interest in antibody discovery and highlight how this knowledge is impacting the antibody engineering field.

THE HUMAN ANTIBODY REPERTOIRE

The primary repertoire of antibodies is produced via the combinatorial rearrangement of IG (H, K, or L)V with IGHD (only in V_H) and IG (H, K, or L)J germline genes, followed by pairing of V_H and V_L domains (Tonegawa, 1983). This repertoire should be diverse and versatile enough to recognize any antigen with a low or medium affinity during the primary immune response (Neuberger and Milstein, 1995). The physical maps of the human IGH and IGL gene loci were elucidated in the 1990s (Tomlinson et al., 1992; Schäble et al., 1994; Matsuda et al., 1998) and the information has been compiled and annotated at The ImmunoGenetics Database (IMGT; <http://www.imgt.org/>). This information has provided the foundations to understand the mechanisms of generation of diversity in human antibodies and has shed light on the evolution of the antibody repertoire.

Overall, the human IGK locus contains approximately 30 functional IGKV genes distributed in six families, and five IGKJ

segments which recombine to form the primary V_k repertoire. There are 30–36 functional IGLV genes arranged in three distinct clusters containing 11 IGLV gene families and four functional C λ domains, each with its own IGLJ gene. The IGH locus contains approximately 39 functional IGHV genes distributed in seven IGHV gene families, approximately 30IGHD segments classified also in seven families and six IGHJ genes. As more human germline genes from diverse individuals have been sequenced and studied, an increasing number of alleles have been compiled at IMGT (Lefranc et al., 2005).

Analysis of the antibody genes amplified from diverse sources (Cox et al., 1994; Huang et al., 1996; Ignatovich et al., 1997; Brezinschek et al., 1998; de Wildt et al., 1999; Farner et al., 1999; Glanville et al., 2009) indicates a strong bias in gene usage. For instance, only five IGHV genes (5–51, 1–69, 1–2, 4–59/61, and 3–30/33) make 50% of the rearranged antibodies and only 24 out of 39 functional genes (~60%) are expressed with a frequency above 1% (Glanville et al., 2009). For IGVK the bias is more dramatic. Only three IGVK genes (3–20, 1–39, and 3–15) make 50% of the rearranged antibody repertoire and only 16 out of 30 genes are expressed with a frequency of more than 1%. Pairing of heavy and light chains in B cells (de Wildt et al., 1999) and in recombinant libraries (Glanville et al., 2009) appears to be a random process, reflecting the relative abundance of the IGHV and IGLV gene family members. The bias in the gene usage is due to a number of factors including position in the locus, ontogenetic regulation of the immune response, gene copy and binding properties of the antibodies encoded by certain genes (Dal-Bo et al., 2011; Lerner, 2011; Zhu et al., 2011).

After antibody exposure to antigen, an affinity maturation process generates diversity from which antibodies with higher affinity are selected, as the antigen concentration decreases during the secondary immune response. Affinity maturation mechanisms include somatic hypermutation (in most mammalian systems) and gene conversion (in certain species, see below). The somatic hypermutation process takes place in the germinal centers with the help of T-cells. The V-genes in activated B cells undergo activation-induced (cytidine) deaminase (AID)-catalyzed somatic hypermutation at a rate of up to 10^{-3} changes per base pair per cell cycle (Rajewsky et al., 1987). Two separate mechanisms are involved in the mutation process (Maizels, 2005); one targets mutation hotspots with the RGYW (R = purine, Y = pyrimidine, W = A or T) motif (Dörner et al., 1998) which includes the reverse complement of the preferential substrate site for AID, while the second incorporates an error-prone DNA synthesis that can lead to a nucleotide mismatch between the original template and the mutated DNA strand (Rada et al., 1998). The overall process favors single-base transitions over transversions at a 3:1 ratio (Betz et al., 1993).

The frequency of mutations in V_H and V_L are qualitatively similar, following an exponential distribution with as much as 15–20% of the V-regions showing no mutations at the amino acid level (Tomlinson et al., 1996; Ramirez-Benitez and Almagro, 2001). The average number of mutations per V-region has been estimated for humans and mice to be around 8 and 5 mutations for V_H and V_L , respectively (Tomlinson et al., 1996; Ramirez-Benitez and Almagro, 2001; Clark et al., 2006). Although the

mutations are spread throughout the V-domains, they occur at a proportion of 3:2:1 mutations at the antigen-binding site, surface of the V-domains and V_L : V_H interface, and core of the V-domain, respectively (Clark et al., 2006). The relatively high proportion of mutations in the CDRs with respect to FRs is explained in part by a higher concentration of mutation hotspots in the former. It also reflects the selection for affinity improvement, although it has been found that somatic mutations in residues in direct contact with antigen are less frequent than in residues adjacent to the residues in contact (Ramirez-Benitez and Almagro, 2001), suggesting that the residues selected during the primary immune response do not change during the affinity maturation. Insertions and deletions also occur but at a lower rate (Wilson et al., 1998; Zhao and Lu, 2010), implying that the overall geometry of the antigen-binding site as defined by the canonical structures does not change significantly during the affinity maturation process either.

The amino acid content and length distribution of the CDR-H3 region is of critical importance to antibody repertoire function and the diversity encoded in this loop in humans has been extensively characterized to aid synthetic mimicry of human diversity (Schroeder et al., 1987; Zemlin et al., 2003; Schroeder, 2006; Glanville et al., 2009). These studies have shown very clearly that the human CDR-H3 repertoire is distinctly different from that of the mouse, particularly in length distribution. The human loops tend to be significantly longer, at 15.2 (± 4.1) residues, while mice average at only 11.5 (± 2.7) (Zemlin et al., 2003). Both species exhibit common conserved motifs at the stem of the loop, but the biases in amino acids used overall and, indeed, in a positional sense, show distinct differences. While humans and mice both show a strong preference for the use of Y, S, and G residues, this phenomenon is much more pronounced in mice (26% Y), than humans (14% Y) and in both species this trend toward high Y use increases proportionally with loop length. In humans in particular, longer CDR-H3 loops are associated with increased use of the IGHJ-6 segment, which encodes a series of contiguous Y residues, increasing the frequency of Y content overall (Prassler et al., 2011; Zhai et al., 2011). In addition, humans use more P and do not exhibit the clear hallmarks of hydrogen bond ladder formation in the loop as often as mice, suggesting more complex overall loop topology in humans. This phenomenon may be directly correlated with increased length in human CDR-H3, with an associated higher use of cysteine via germline-encoded “D2” DH sequences. These long D-segments encode for cysteine residues spaced four amino acids apart, allowing disulphide loop formation that can be critical to CDR secondary structure and rigidity (Almagro et al., 2012). While these disulphide-stabilized loops are relatively rare in humans (C = 1.21% of all amino acid use in human CDR-H3), (Zemlin et al., 2003) they are a commonly used motif in the antibodies of both chickens and camelids, as outlined later.

HARNESSING NON-HUMAN ANTIBODY V-GENE REPERTOIRES

Species such as mouse, chicken and camelids (such as llama) are all used as immune sources of antibodies with therapeutic potential. While antibodies from human libraries theoretically contain “fully human” amino acid sequence in their FRs, antibodies from

immune animal repertoires do not. Nonhuman-derived antibodies may initially have their immunogenicity reduced by cloning the V-genes onto a set of human C regions, to form a “chimeric” antibody (Morrison et al., 1984). Even the small amount of “foreign” amino acid content with respect to humans in the V-domain FRs of a chimeric IgG may be enough to provoke an anti-idiotypic antibody response, however, especially patients that receive repeated doses of antibody as therapy (Stephens et al., 1995). As a result, before clinical use, antibodies derived from animals usually undergo a process of “humanization,” whereby recombinant DNA technology is used to “graft” the CDRs of the clone of interest onto human V-gene framework scaffolds (Jones et al., 1986). It is typically necessary to carry out subsequent V-gene engineering, e.g., via “back mutations” in the FRs, to return the target binding affinity of the parental clone (Almagro and Fransson, 2008). For this humanization process to be efficient, it is helpful not only to be able to predict which human FRs might be optimal to accept the grafted CDRs from a lead clone, but also to understand the nuances of the structural characteristics of the repertoire from which the clone was derived. Armed with sufficient prior knowledge of each species’ repertoires, we can confidently predict the likely engineering path that will be required to derive a fully active, but maximally humanized product.

THE MOUSE ANTIBODY REPERTOIRE

The mouse (*Mus musculus*) is the most widely used model organism in immunology and perhaps in biology and medicine. For the study of antibodies, the development of hybridoma technology, first described by Köhler and Milstein (1975) and awarded the Nobel Prize in 1984, was the key advancement that ultimately led to development of antibody-based drugs. Hybridoma technology involves the immunization of rodents with an antigen of interest and once a satisfactory immune response against the antigen has been obtained, the antibody-producing B cells are harvested and fused to a murine myeloma cell line. The resulting hybrid cells can be sub-cloned to generate clonal cell lines in which every cell secretes antibodies with a single specificity. Thus, hybridoma technology became an efficient means to produce unlimited amounts of single-specificity antibodies which enabled the biochemical and structural characterization of antibodies and the production of sufficient quantities of high quality protein for therapeutic settings.

The physical maps of the mouse IGH and IGL gene loci were elucidated in the second half of the 1990s (Tomlinson et al., 1992; Schäble et al., 1994, 1999; Matsuda et al., 1998; Thiebe et al., 1999) and, as for humans, the information is compiled and annotated at IMGT (<http://www.imgt.org/>). The total number of mouse (*M. musculus*) IGK genes per haploid genome is 164 (174 if the orphans are included), of which 99 are functional, belonging to 18 subgroups (Martinez-Jean et al., 2001). Eighty-one are in opposite orientation of transcription, 59 of them are functional and must rearrange by a mechanism of inversion. These genes are recombined with five IGKJ genes. The IGL locus contains only three IGLV genes each with one associated IGLJ gene. The reduced contribution of the IGL locus to the mouse germline repertoire is consistent with the approximately 8-fold reduction

in the prevalence of lambda-bearing IgG in the serum of mice compared to humans.

The IGH locus is both larger and more diverse than that of the humans (Schroeder, 2006). IMGT reported as of August 2012 two tables for the mouse IGVH germline gene repertoire. One with IGHV genes compiled from diverse sources, which represent genes characterized in several strains and thus some genes may be alleles. The other table compiles data from the C57BL/6 Mouse Genome Sequencing and is provisional since not all the genes have been mapped and confirmed. It lists 170 IGVH germline genes distributed in 15 IGHV gene families. One hundred one out of the one hundred seventy known genes (~60%) are functional genes. These IGHV genes recombine with 21 functional IGDH genes assorted in four families and four IGHJ functional genes.

Interestingly enough, comparisons of the canonical structure repertoire encoded in mouse and humans IGHV genes (Almagro et al., 1997; Bono et al., 2004) indicate that the human structural repertoire has two additional classes (1–1 and 1–3). Thus, the human repertoire is more diverse in structural terms than that of mouse. In addition, the canonical structure class 1–2 is more prevalent in mouse (~60%), while in humans the dominant class is 1–3 (~40%) (Almagro et al., 1997). This divergence, together with phylogenetic analysis of the human and mouse IGHV genes (Bono et al., 2004), indicates that most of the sequences in the human and mouse IGHV loci have arisen subsequent to the divergence of the two organisms from their common ancestor. Identifying these differences between human and mouse genes, which are perhaps a reflection of functional and/or structural constraints at work to balance the free diversification of the antibody repertoire in humans and mice (Almagro et al., 1997), could be useful to select the most human-like genes for humanization of mouse antibodies.

THE CHICKEN REPERTOIRE

Gallus gallus, the domestic chicken, is a classic model for immunological study. Indeed, “B-cell” derives from the term “Bursal cell,” as B-cells were first recognized as products of the Bursa of Fabricius, a cloaca-associated organ that is critical to immune development in birds (Ratcliffe, 2006). The antibody repertoire of chickens has also been extensively characterized in functional isotype content and at the genomic level (Reynaud et al., 1985, 1989; Ratcliffe, 2006). Their immunoglobulin system is distinct from that of humans and mice as they have structural equivalents of mammalian IgM, IgA, and IgG, but not IgE or IgD. IgM is the major isotype expressed on the surface of their B-cells (Ratcliffe, 2006). Additionally, all chicken antibodies use λ isotype light chains, exclusively (Reynaud et al., 1987). Chicken IgG has 4 Cy domains, however, and is thought to be a structural relative of both mammalian IgG and IgE subclasses (Parvari et al., 1988). Chicken IgG is also found in a “short” form, lacking the CH3 and CH4 regions. Avian IgG is often described as “IgY” as it can be found at high concentration in egg yolk, but it has been proposed that the full-length form should be called IgG and the short form IgY, to aid their differentiation (Ratcliffe, 2006).

The avian V-gene germline repertoire is extremely simple, with single functional V-genes in both the light and heavy chains,

that contain unique V_L - J_L and V_H - D - J_H segments (Reynaud et al., 1989, 1991; Parvari et al., 1990). The chicken V_L and V_H germline domains are highly homologous to stable and soluble human V_{λ} and V_H families (Ewert et al., 2002, 2003), respectively, and this is maintained across the fully mature repertoire (Wu et al., 2011). The uniformity of chicken V-gene FW sequences renders them highly predictable in humanization (Tsurushita et al., 2004; Nishibori et al., 2006). Despite this simple V-gene system (Reynaud et al., 1983; Parvari et al., 1987a,b, 1988; Ratcliffe, 2006), chickens have a broadly adaptable repertoire that generates high affinity antibodies to protein, peptide and haptens antigens (Yamanaka et al., 1996; Finlay et al., 2005, 2006; Nishibori et al., 2006).

The chicken V-gene system is in stark contrast to that found in humans, mice and primates, which all utilize a large set of V-gene sequences that are highly diverse in both sequence and structure (Schroeder et al., 1990; Schroeder, 2006). In chickens, as in rabbits (Weill and Reynaud, 1992), a distinctly different set of diversification mechanisms are used, including gene conversion (Reynaud et al., 1987, 1989). Gene conversion relies on a single template V-gene being diversified via the incorporation of segments from upstream pseudogenes that lack recombination signal sequences. This process is used to diversify both the heavy and light chains, with mutations being introduced into both CDRs and FRs. For the process to be efficient, it relies on high sequence homology between the pseudogene and the germline gene which acts as the acceptor (Ratcliffe, 2006). A recent chicken V_H repertoire analysis suggests the requirement for sequence homology between germline and pseudogene leads to a low level of mutagenesis in the FWs, but hypervariability in the CDRs (Wu et al., 2011). Interestingly, this was coupled with strong maintenance of common CDR structural residues that have also been observed in mammals (Rader et al., 2000; Zemlin et al., 2003; Lee et al., 2004), but modulation of residues that affect V_H - V_L interaction (Padlan, 1994) and CDR structure (Foote and Winter, 1992). The chicken V_H repertoire therefore adds significant variability at select FW positions to increase structural diversity, e.g., by changing the angle of interaction between the V_H and V_L domains (Abhinandan and Martin, 2010).

The CDR-H3 repertoire of chickens differs distinctly from that of humans and mice, in both length distribution and amino acid content. Surprisingly, chickens have only 15 functional D-segments, all of which are highly homologous and some (e.g., D9/12/13, plus D4/8/11) are even identical in amino acid sequence (Reynaud et al., 1991). Additionally, reading frame 1 predominates in chickens (Raaphorst et al., 1997), as reading frames 2 and 3 create sequences containing stretches of hydrophobic residues and stop codons, respectively (Reynaud et al., 1991; Weill and Reynaud, 1992). This form of reading frame control appears to be universal and has also been observed (albeit in different reading frames) for; rabbits, sharks, mice, primates, and humans (Raaphorst et al., 1997; Schroeder et al., 1998; Schroeder, 2006). In reading frame 1, chicken D-segments are biased toward the use of G, S, and Y, as observed in all other vertebrate species studied to date (Zemlin et al., 2003; Schroeder, 2006). In contrast to humans and mice however, chicken D-segments obligately contain C, with the consensus

sequence G-S- (A/G)-Y-C- (G/C)- (S/W)-X-A- (Y/E) (X = non-conserved) (Reynaud et al., 1991). This limited initial V_H CDR3 repertoire is hyper-diversified both by somatic mutation and the insertion of new sequences via gene conversion. These D-like sequences are donated by pseudogenes and may replace the entire D-segment or only a small section, leading to the creation of “mosaic CDRs” (Reynaud et al., 1989, 1991).

Analyses of CDR-H3 amino acid content in the chicken shows very different paratope chemical composition in comparison to humans and mice (Wu et al., 2011). There is a distinct bias toward small amino acids G/S/A/C/T (but not P), while large aromatic and hydrophobic residues are strongly disfavored, including an unusually low representation of Y, the dominant residue in the repertoires of mice and humans (Zemlin et al., 2003). This observation may be important, as synthetic antibody repertoire studies have suggested that Y is a critical amino acid for target binding (Fellouse et al., 2004, 2005, 2006, 2007). Additionally, the chicken CDR3 repertoire has low representation of positively charged residues (K/R). This may be of practical importance, as excess positive charge in the V_H CDR3 is associated with polyreactivity (Li et al., 2001) and poor pK profile *in vivo* (Boswell et al., 2010).

The use of C in the CDR-H3 of >50% of all B-cell clones in the chicken repertoire is suggestive that it plays an important functional role. While humans and rhesus make functional CDR-H3 sequences containing a pair of cysteines (Zemlin et al., 2003; Schroeder, 2006), these are found at low frequency in mature human B-cells, and they are very rare in mice (Raaphorst et al., 1997; Zemlin et al., 2003). The high incorporation rate of C in the chicken CDR-H3 is rendered functional by two mechanisms: (1) frequent use of D-D junctions (Reynaud et al., 1991) to create CDR3s with intra-CDR disulphide bridges and (2) insertion of single C residues in the V_H CDRs 1 and 2 for inter-CDR disulphide bonding. These covalent bonds between CDRs are structurally analogous to those observed at high frequency in the immunoglobulins of other species such as camelids (Harmsen et al., 2000), sharks (Dooley et al., 2003; Stanfield et al., 2004), cows (Aitken et al., 1997; Sinclair et al., 1997; O'Brien et al., 1999), pigs (Li and Aitken, 2004), and even the duckbilled platypus (Johansson et al., 2002). It seems likely that the increased use of disulphide binding in long CDRs, by several species, may be highly beneficial to stabilize longer loops that have greater sequence diversity, but could suffer from a lack of structural rigidity that leads to an entropic penalty during binding interactions (Wong et al., 2011; Hackel et al., 2010). Mutagenesis studies have shown that these disulphides, in either IgG or single-domain antibodies, are essential for both V-domain stability and binding function (Lee et al., 2006; Fennell et al., 2010; Govaert et al., 2012).

BEYOND STANDARD IgG STRUCTURES—NATURAL ‘DOMAIN ANTIBODIES’

Despite being the main format for many successful therapeutics, IgG molecules have some practical limitations as they are large (~150 kDa), covalently-linked tetrameric structures that classically contain two antigen-binding sites. The necessity for two V-regions to combine and stabilize each other makes it technically challenging to reduce antibodies to anything smaller

than the dual-domain single chain Fv (scFv) antibody fragment (~30 kDa). The desire for smaller, more stable and monomeric binding modalities in appropriate indications has led to the investigation of a logical alternative; modular therapeutics built from naturally-occurring binding proteins that can be used as a source of “domain antibodies.” As outlined below, comparative immunogenetics led to the discovery of non-classical immune proteins such as the camelid VHH and the shark VNAR (variable domain of the IgNAR), which can both be isolated as soluble, stable, monomeric V-domains (**Figure 6**) (Flajnik and Dooley, 2009; Wesolowski et al., 2009; Flajnik et al., 2011). These single domain proteins are only ~12–15 kDa in size and have been the subject of significant academic and industrial research to characterize their origins and utilities (Muyldermans et al., 2009; Flajnik et al., 2011). Humanization of VHH antibodies is facile, as the isolated antibodies are typically close to a human VH germline sequence. Together with high stability and low aggregation, this gives the humanized VHH antibody theoretically low immunogenicity risk. To date, the less heavily investigated IgNAR has not been extensively characterized in humanization studies and may represent a different challenge from VHH. The VNAR domain is actually more structurally related to a T-cell receptor α -domain and has much lower a.a. identity to human homologous domains. As a result, in this section we concentrate on the more experimentally advanced VHH.

Domain antibodies, lacking an Fc region, suffer from fast renal clearance and without protein engineering they have short *in vivo* half-lives. Luckily, the stable, soluble nature of isolated domain antibodies renders them relatively simple to engineer in a modular fashion. Modifications at the N- or C-terminus are typically possible without loss of function, allowing fusion to common half-life extension molecules such as serum albumin

or immunoglobulin Fc and covalent conjugation to natural or non-natural polymers that expand the hydrodynamic radius of the proteins, greatly reducing renal clearance. Domain antibodies have also been extensively exploited as modular units to create bispecific molecules for targeting multiple disease mediators with a single polypeptide (Gill and Damle, 2006; Harmsen and De Haard, 2007). A final exciting avenue open to domain antibodies is the possibility of oral administration, e.g., via strains of *Lactobacillus* expressing a TNF-specific VHH antibody, which was efficacious in a murine gut inflammation model (Vandenbroucke et al., 2010). In the following sections we outline the current state of knowledge surrounding these molecules and the influence that combined repertoire and structural analyses have had on their understanding and application.

THE CAMEL ANTIBODY REPERTOIRE AND VHH

In 1993, the Hamers’ group identified a previously unknown immunoglobulin form observed in camel serum (Hamers-Casterman et al., 1993). This new immunoglobulin was found not only to lack a light chain, but to have also deleted the CH1 domain in the heavy chain, following the loss of the splice consensus site (Nguyen et al., 1999). These unique, “heavy chain antibodies” were shown to have a VHH-Hinge-CH2-CH3 structure and performed their antigen binding function exclusively via a stable and soluble VH domain, subsequently dubbed “VHH.” VHH antibodies were found to be fully functional constituents of the immune repertoire of camels, representing >50% of the total Ig population in serum samples (Muyldermans and Lauwerys, 1999; Nguyen et al., 2001). Later studies have shown that other camelids such as llama and alpaca also share these unusual immunoglobulins (Harmsen et al., 2000). The VHH and IgG camel antibodies can be separated by isotype, with IgG1 using architecture of conventional antibodies, whereas the IgG2 and IgG3 isotypes are associated with VHH antibodies (Flajnik et al., 2011).

While they are simple in structure, immunogenetics studies have shown that rather than being a rudimentary evolutionary form of immunoglobulin, the VHH antibody in camelids was actually derived from the genes of a conventional IgH locus by a relatively recent adaptation (Nguyen et al., 2002). Multiple contributory selection pressures have been postulated that might have driven this evolutionary event including; amyloidosis associated with a key light-chain sequence, a virus that targeted a light-chain as a co-receptor, or a simple biophysical pressure to develop high frequency antibodies with a “protruding” CDR structure that is highly appropriate for probing cryptic epitopes (Flajnik et al., 2011). Indeed, multiple co-crystal structures of both VHH (De Genst et al., 2006) and IgNAR (Stanfield et al., 2004, 2007) in complex with enzymes have shown the CDRs to protrude into the active site cleft of the enzyme, neutralizing its function.

Despite relying on a single variable domain for antigen recognition, it has been shown that the VHH repertoire is as complex in sequence diversity as its VH counterpart in camelid IgG1 (De Genst et al., 2006). Indeed, although camelid VHH and VH domains are encoded by distinct sets of V-gene segments, both forms of antibody share some D segments and an identical JH region (Nguyen et al., 2001). Similar to chickens,

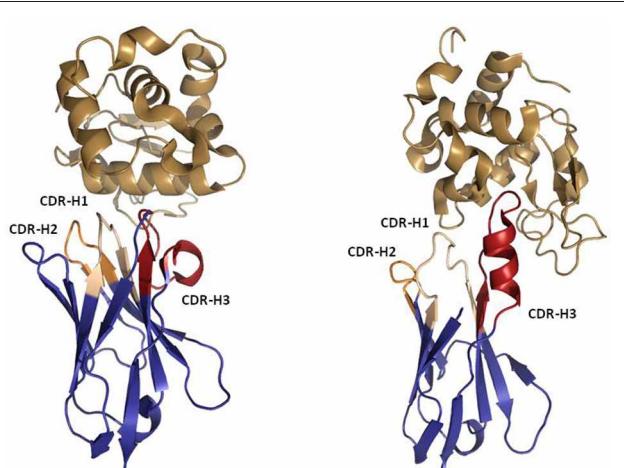


FIGURE 6 | Ribbon representation of Llama VHH (left) and Shark IgNAR (right) in complex with Hen Egg White Lysozyme (HEL). PDBID 1OP6 was used to represent the VHH:HEL complex. PDBID 2I26 was used to represent the IgNAR:HEL complex. Note the protruding IgNAR CDR-H3 blocking the active site of HEL, which contrasts with the bended VHH CDR-H3 that recognizes a flat epitope. The Figure was generated using PyMOL.

sequence analysis of camelid VHH domains has shown very close homology to the human V_H3 family, which is also associated with relatively high stability and solubility (Ewert et al., 2002, 2003). Comparative analyses of VHH and VH germline and repertoire sequences have shown clearly important differences in their respective structures (Riechmann and Muyldermans, 1999; Harmsen et al., 2000), with VHH exhibiting higher frequency of hypermutation hotspots, leading to greater diversity in CDR-H1 and CDR-H2 sequences and length, plus the frequent observation of clones with long CDRs 1 and 3. In another convergence with chickens, VHH antibodies frequently use non-canonical C residues in their CDRs (Govaert et al., 2012). While the disulphide bonding patterns seen in camelids are not as varied as those observed for chickens (Wu et al., 2011), they do lead to disulphide bonding within the CDR-H3, between CDR-H3 and CDR-H1, or between CDR-H3 and FR-2, with the cysteine groups outside the CDR3 typically being placed in very similar positions to those observed in chickens (IMGT positions 38, 55) (Harmsen et al., 2000).

Most critical of all known VHH characteristics, are the hydrophobic to hydrophilic substitutions of four critical residues in the FR-2, known as the “VHH tetrad.” These residues in the FW2 are in critical positions where the V_H of a conventional IgG would pack against the V_L (Abhinandan and Martin, 2010), providing hydrophobic binding affinity between the two domains. The classic substitutions V37F/Y, G44E, L45R, and W47G lead to a major increase in hydrophilicity of the VHH, allowing it to fold and function independently, without the need for a stabilizing V_L partner (Harmsen and De Haard, 2007). This adaptation is essential for biotechnological use, as it allows expression of VHH antibodies at high concentration. While some conventional V_H domains can be expressed, they will typically become insoluble at concentrations above 1 mg/ml (Davies and Riechmann, 1994).

Importantly, while long CDR-H3 loops may be common for some VHH sub-types (particularly in camels) and can contribute to solubility by folding over the FR-2, in llama VHH the average CDR-H3 length has been shown to not exceed that observed for humans (Harmsen et al., 2000). Indeed, experimental analyses of independent V_H domains derived from chicken IgGs, which do use long CDRs (Wu et al., 2011) but do not contain the FR-2 “tetrad” substitutions, have shown that these domains do not exhibit high solubility and do lose binding affinity when separated from a light chain partner (Finlay et al., unpublished observations). Long CDR-H3 sequences are therefore not a guarantee of FR-2 coverage or solubility in VHH and the FR-2 tetrad appears to be an essential factor in achieving solubility. Studies on the “camelization” of VH domains isolated from monoclonal IgG antibodies have corroborated this, by showing that the addition of the FR-2 tetrad mutations can significantly improve the solubility of those domains (Davies and Riechmann, 1994; Riechmann and Muyldermans, 1999). Nonetheless, camelized and/or CDR-solubilized human antibody domains struggle to replicate the qualities of natural VHH, (Barthelemy et al., 2008) suggesting that combined FR and CDR repertoire content may also play a major role. At the time of writing, no major repertoire analyses have been performed for VHH in the way they have for humans, mice and chickens (Zemlin et al., 2003; Wu et al., 2011).

MAN-MADE ANTIBODY REPERTOIRES

Phage display technology was developed by George Smith in 1985 (Smith, 1985) to display peptides on the surface of the filamentous bacteriophage M13. In an effort to isolate “fully human” antibodies and thus bypass humanization, phage display was adapted at the beginning of 1990s (McCafferty et al., 1990) to display antibody V-domain repertoires and to isolate antibodies of interest *in vitro*. During the 1990s and the last decade, several academic laboratories and biotechnology companies have designed and implemented human antibody phage-displayed libraries for antibody discovery (Hoogenboom, 2005; Bradbury, 2010). Such libraries have enabled the isolation of high affinity and specific antibodies against a wide range of molecules and the antibody library design and implementation process continues to evolve.

Since phage display bypasses immunization, it is especially useful for obtaining antibodies against targets that are highly conserved across species and those that may be toxic, where *in vivo* methods are ineffective and/or impractical. In addition, since phage display technology allows access to the repertoire of genes intended for expression and display on the phage surface, the number of genes and variants can be designed or chosen to bias the repertoire toward genes with predefined characteristics, opening up the possibility of testing hypotheses on how the size of a repertoire, its diversity and composition impact the selection of specific, stable, soluble, and high affinity antibodies. Overviews of different man-made repertoires and how their performance has enhanced our knowledge of the evolution of the antibody repertoire are provided below.

NATURAL (NAÏVE) REPERTOIRES

Originally, human antibody phage-displayed libraries for *in vitro* discovery were implemented either by cloning the natural repertoire of rearranged antibody genes (Marks et al., 1991) or by rearranging human antibody germ-line genes *in vitro* (Griffiths et al., 1994). This first generation of natural or naïve repertoires contained the diversity harvested from the total B-cell repertoire by RT-PCR and thus suffered from a lack of control over FR usage and mutation rate (Sidhu and Fellouse, 2006). This can be a significant issue in antibody therapeutic development, as not all human FRs are used at high frequency in the B-cell repertoire (see above) and, importantly, not all will express well in heterologous systems or be stable in delivery formulations (Ewert et al., 2003). Additionally, antibodies from naïve libraries may contain somatic mutations leading to FR or CDR based T-cell epitopes and/or aggregation-prone sequences that could potentially lead to immunogenicity (Harding et al., 2010). As a result, several research groups have developed fully synthetic antibody repertoires in which a few well-expressed and well-behaved scaffolds are used for the repertoire synthesis and diversity is restricted to the CDRs (Pini et al., 1998; Sidhu et al., 2004).

SYNTHETIC REPERTOIRES

The earliest experimental synthetic antibody repertoire was based on single V_H and V_L scaffolds, introduced limited diversity into the CDR-H3 alone and was used to generate moderate affinity hits against haptens (Barbas et al., 1992, 1993). This library exploited basic knowledge of antibody structure and diversity by

placing random amino acid diversity into the exposed regions of the CDR-H3. Amino acid randomization was made possible by the use of PCR and oligonucleotides containing degenerate DNA sequence in the appropriate CDR-encoding codons. The simplest forms of degenerate codons used are based on random incorporation in the first two bases, followed by restricted incorporation in the third position to either G/T ("NNK" codon) or G/C ("NNS" codon). Both of these schemes predominantly incorporate functional diversity, as they encode for 32 codons total, including only one of the three stop codons (amber) and encoding all 20 amino acids, although not at equal ratio. These codons therefore incorporate a reasonable number of translatable polypeptide sequences, so long as the number of contiguous codons used is not so many that one stop codon would be expected per clone. One complicating factor is the obligate encoding of cysteine by each of these codons, leading to a significant number of clones in which thiol groups are presented unpaired, leading to loop malfunction and reducing the overall functional content of the library.

This very simple method of diversification of key CDR loops was exploited by a series of teams (Griffiths et al., 1993, 1994; Viti et al., 2000; Silacci et al., 2005), who used similar methods to diversify both the CDRs H3 and L3 in a variety of FRs, while maintaining key loop stem residues in the CDR-H3 such as Kabat 93, 94, 101, and 102. Despite the simple nature of these repertoires, they have been highly successful at generating antibodies to proteins, peptides, and haptens. The isolated antibodies have proven utility as immunochemical reagents (Neri et al., 1998) and some have even been applied successfully in therapeutic settings (Neri et al., 1995; Carnemolla et al., 1996; Borsi et al., 1998; Brack et al., 2006; Silacci et al., 2006). This antibody construction style was then progressed by examining the additional benefit of adding synthetic diversity in the CDRs 1 and 2 of V_H and exploiting the use of a variety of tailored (A.K.A. "parsimonious") degenerate codons that encoded a smaller number of amino acids at key positions. This improved the quality of the resulting libraries, so they encoded a larger proportion of functionally folded clones and the avoidance of stop codons allowed the incorporation of greater length diversity in the CDR-H3 (Lee et al., 2004; Sidhu et al., 2004).

Further examination of the amino acid content of the mature human, primate and rodent V-gene repertoires, coupled with expanded structural understanding of antibody-antigen interactions, as outlined above, subsequently created highly defined knowledge of positional amino acid usage in CDRs. These studies strongly suggested that the key amino acids used in making functional contacts with protein antigens were highly biased and the adoption of TRInucleotide, or "TRIM" technology (Virnekas et al., 1994) allowed the first opportunities for this knowledge to be fully exploited. TRIM technology is a method based on classical oligonucleotide synthesis chemistry, but at positions of randomization, mixes of trinucleotide phosphoramidites (A.K.A. trimers) are added instead of a series of single base mixes. Each trimer is a fully synthetic codon and the use of precise combinations of these trimers therefore allows the incorporation of positional bias in amino acid mutagenesis libraries (Virnekas et al., 1994).

The earliest use of TRIM technology in antibody repertoire construction did not exploit its full capability, only introducing fully random amino acid diversity into the CDR-H3 of a single framework pair and selecting the library successfully against a series of haptens (Braunagel and Little, 1997). Subsequent landmark studies, however, took the use of trinucleotides to its logical conclusion and attempted to closely mimic the natural human immune repertoire in synthetic form (Knappik et al., 2000; Rothe et al., 2008; Prassler et al., 2011). Knappik et al. (2000) generated the first iteration of the HuCal[®] libraries, in which they first recognized that 95% of all human antibody diversity is represented in only seven V_H and seven V_L germline gene families. This observation inspired them to create a library of V-genes built on consensus FRs derived via alignment of each of these families. Into these FRs they placed double-stranded CDR diversity "cassettes" that had been built using trinucleotides to represent each of the amino acids naturally found at all positions in CDRs H3 and L3. They also included length diversity in the CDR3s and canonical structural determinants for the L3, approximately mimicking the natural biases observed in the human repertoire. In the CDR-H3, the natural dominance of G and Y in the human repertoire was closely reflected, with those two residues making up approximately 15% each, of the encoded residues between Kabat 95 and 100s. All other residues were included at ~4% other than cysteine, which was allowed at only ~1% to allow potential generation of the disulphide-constrained loops that are occasionally observed in the human repertoire. Stem loop biases in both the CDRs H3 and L3 were also closely maintained and limited diversity was introduced at several key positions in the structurally conserved V_{κ} and V_{λ} L3 loops. This repertoire design was highly effective in generating nM-affinity clones with specificity for a selection of proteins and peptides (Knappik et al., 2000; Marget et al., 2000).

The methods used by Knappik et al. (2000) have since been elaborated upon in a number of reports. Rothe et al. (2008), reported the construction of the next-generation library HuCal Gold[™] in which the design strategy was refined to change the CDR design complexity, the structural format of the library (to Fab fragment) and also to move to "CysDisplay" in which the expressed antibody is tethered to the phage via a disulphide bond with a mutant p3 protein. In this library, the CDR diversity was extended to the CDRs 1 and 2 of both V-domains, in addition to the CDR3s. The CDR cassettes were again based on trinucleotide technology and were built to accurately reflect the canonical structures and diversity found in the families upon which each of the consensus frameworks were built, consciously including structures known to be preferred in the recognition of peptides. The finalized library was found to be of very high practical utility, routinely generating antibody specificities and affinities in the single digit nM range that were useful for both therapeutic and reagent purposes (Jarutat et al., 2006, 2007; Ohara et al., 2006; Prassler et al., 2009). This library design had, however, made a critical concession to optimize its function in *E. coli* expression and phage display technology: the FRs were codon optimized specifically to maximize for *E. coli* periplasmic expression rate. In addition, this function had been aggressively selected for by pre-screening all antibody sequences for periplasmic transport

as β -lactamase fusions before inclusion in the final library. This resulted in a high frequency of antibodies being selected that performed very well in prokaryotic expression, but poorly in mammalian cell lines used for industrial production of IgGs. As a result, a third iteration “HuCal Platinum™” has since been made which further updated the design and performance of these synthetic libraries, by optimizing codon use to suit both prokaryotic and eukaryotic expression systems (Prassler et al., 2011). This library also further refined the FR use and CDR content to minimize T-cell epitope content and maximize similarity to the human repertoire, by adding length-dependent positional amino acid bias in the CDR-H3 and by switching some V_H gene families to fully germline (e.g., VH3-23). These changes made the library higher performing than the HuCal Gold™ library in diversity of hits generated, average affinity and expression rate in mammalian cells. Similar libraries that have recently been generated by separate groups have also strongly supported the broad utility and quality of libraries that naturally mimic the human immune system and that these libraries are flexible in display format as they can be selected by alternative phage display systems such as pIX display (Shi et al., 2010).

The overall HuCal™ story is therefore a clear paradigm and example of the intrinsic attraction of synthetic antibody libraries: they can be designed to add positive attributes and to minimize negatives. While fully human natural cDNA-derived libraries are simple to construct and highly functional, negative attributes such as unwanted FR use, somatic hypermutation in frameworks and potential liability sequences such as aggregation motifs, deamidation sites, N-linked glycosylation motifs, oxidation sensitivities and non-canonical disulphide content cannot be avoided as a rule. The very latest libraries make use of a novel DNA synthesis technology known as Slonomics, which has been shown to be an excellent method for the synthesis of molecular diversity, as it allows the production of precise amino acid/codon biases and low dysfunctional sequence content at any given position (Zhai et al., 2011). Analysis of the first large library generated using this technology strongly supports another benefit of carefully designed synthetic antibody libraries: the removal of segmental linkage within the CDR-H3 that limits paratope structural diversity. While natural CDR-H3 sequences have the benefit of being selected for function in the B-cell, synthetic versions are not generated by IGV-D-IGJ recombination mechanisms and therefore, escape potential limitations on self-reactivity that may be imposed by natural tolerance mechanisms (Zhai et al., 2011). Such synthetic antibody libraries are clearly a mature technology and have become an important section of the armamentarium currently in use for therapeutic human antibody discovery.

MINIMALIST REPERTOIRES

A series of illuminating studies set out to examine the validity of the suggestion that only certain amino acids, with particular emphasis on Y (see above), are essential in the formation of a functional antibody repertoire (Fellouse et al., 2004, 2005, 2006, 2007; Birtalan et al., 2008, 2010; Fisher et al., 2010). It was postulated that antibody critical contacts were predominantly mediated by Y and that small amino acids played a critical support role in creating conformational flexibility and diversity, especially in the

CDR-H3 (Koide and Sidhu, 2009). To examine this hypothesis, large phage libraries of human Fabs on a single FR combination were synthesized containing only four amino acids; Y, A, D, and S in solvent-exposed CDR positions, including CDR-H3 (Fellouse et al., 2004). These libraries were capable of generating high affinity and specific antibodies, including antibodies of 2 nM affinities for VEGF. Remarkably, when the four amino acid code was reduced to only Y and S, highly functional repertoires could still be generated, (Fellouse et al., 2005, 2007) with structural studies subsequently showing that Y was indeed the key residue for making critical contacts with antigen, while serine predominantly provided structural flexibility and “space” to accommodate the bulky Y side-chains (Fellouse et al., 2006).

Importantly, Birtalan et al. (2010) have since progressed these studies and demonstrated that W is the only natural amino acid that can be used as a functional alternative to Y in experimental synthetic antibody libraries based on binary diversity (Birtalan et al., 2010). Indeed, co-crystal structure analysis of an antibody from a W/S-containing repertoire with its target HER2 showed clearly that W is a key determinant of the binding specificity (Fisher et al., 2010). Additionally, these investigators have shown that excess content of highly charged residues such as R is a significant risk factor for the high frequency generation of polyreactive clones (Birtalan et al., 2008, 2010). A subsequent study performed using a set of libraries of synthetic single-domain binding proteins has provided supporting evidence for the idea that minimal diversity can be genuinely functional, but that increased amino acid diversity overall is still the best for high function (Hackel and Wittrup, 2010). Synthetic antibody libraries have, therefore, not only allowed the interrogation of fundamental antibody structure/function relationships, but have illustrated what library design elements are critical to maximal function. Future studies will most likely continue to advance this field to make synthetic diversity as reliable as possible.

RATIONAL REPERTOIRES

Finally, based on the finding that the anatomy of the antigen-binding site determines the propensity to recognize a defined type of generic antigen such as a peptide or a hapten, it has been hypothesized that by biasing an antibody repertoire toward the recognition of predefined antigens the probability of obtaining more specific and higher affinity antibodies may increase. This can be rationalized in terms that, when general purpose repertoires, such as naïve or synthetic repertoires, are used to obtain antibodies to a given target, a vast region of the shape space has to be explored to produce specific antibodies. Provided that only an infinitesimally small fraction of all possible functional antigen-binding sites can be explored by using enrichment technologies, such an exploration should be sparse. Consequently, the probability of selecting specific antibodies of higher affinity should increase in repertoires that have been designed to be focused on predefined regions of the shape space.

Rational repertoires of antibodies are built by selecting genes encoding combinations of canonical structures that resemble the structural features of antibodies that bind the desired type of ligands. Sequence diversity is then introduced at residues typically involved in recognition of those types of targets. For instance, two

antibody repertoires have been designed and tested for peptide recognition (Cobaugh et al., 2008). First, a human anti-peptide repertoire was constructed by pairing the human IGVH germ line gene 3–23 with a variant of the IGVK germline gene 3–20. The CDR-L1 of the gene 3–20 was modified to encode a long loop, typical of anti-peptide antibodies (**Figure 7**) and diversity was engineered in V_H SDRUs of anti-protein and anti-peptide antibodies (see above). Another repertoire was generated using the V-regions of the murine antibody 26–10, which was originally isolated, based on its affinity to the hapten digoxin, but also binds peptides and exhibits a canonical structure pattern typical of anti-peptide antibodies. As in the first repertoire, diversity was introduced in V_H only, using the profile of amino acid found at positions that frequently contact peptide antigens. Both repertoires yielded binders to two model peptides, angiotensin and neuropeptide Y, following screening by solution phage panning. The repertoire built onto the 26–10 scaffold yielded antibodies with affinities below 20 nM to both targets.

Another example of a rational repertoire (Persson et al., 2006) was designed using as scaffold the antibody FITC8, which has a cavity in the antigen-binding site, common to anti-hapten antibodies. In five CDRs, diversity was designed on the basis of a 3D model structure of FITC8 and anti-hapten SDRUs. In addition, length variation was introduced into the CDR-H2, as longer versions of this loop have been shown to correlate with increased hapten binding. The repertoire was cloned, phage-displayed and screened against a panel of five haptens, yielding diverse and highly specific binders to four of the selectors. Parallel selections

were performed with a repertoire having diversity in more peripherally located residues, which are more often found in contact with protein than haptens. The binders selected from the control (anti-protein) repertoire were not able to bind to the soluble hapten in the absence of the carrier protein, in contrast to the clones selected from the anti-hapten repertoire. Thus, although more validation is needed in order to conclude that rational repertoires increase the probability of obtaining more specific and higher affinity antibodies, these two examples have shown the feasibility and potential advantages of designing repertoires to recognize predefined generic ligands.

CONCLUSIONS AND FUTURE DIRECTIONS

In this review, we have outlined the importance of the continuum of knowledge that runs through antibody structural studies, species repertoire analyses and experimental exploitation of this information to design antibody repertoires and advance antibody-based drug discovery. Study of hundreds of x-ray crystallography antibody structures free and bound to wide variety of ligands have provided a detailed picture of how antibodies recognize diverse types of ligands with exquisite specificity and high affinity. It has also shown that, although the antigen-binding site is very diverse in sequence and structure, it has predictable geometrical features that determine the types of generic ligands with which the antibody interacts. This knowledge has been critical to understand the mechanisms of the immune response mediated by antibodies and the evolution of the antibody repertoire. Its application has led to the development of engineering methods such as humanization, antigen-affinity optimization and effector function enhancement, which have made possible the approval of more than 30 antibody-based medicines in the last two decades.

The knowledge gained on the antibody structure has been complemented with the study of the antibody repertoire of several species. In addition to humans, we described in previous sections the repertoires and the affinity maturation mechanisms of mice and chickens, plus the use of novel single-domain antibodies in camelids and sharks. These species all utilize diverse evolutionary solutions to generate specific and high affinity antibodies and illustrate the plasticity of natural antibody repertoires. Their comparative study has raised fundamental questions about the evolutionary factors shaping the antibody repertoire such as: has the evolution of the antibody repertoire been a stochastic process or has it been shaped by functional and/or structural constraints? What is the optimal size and diversity of a repertoire that can be generated *in vitro* in order to generate specific and high affinity antibodies to a wide variety of antigen types?

Multiple variations of man-made antibody repertoires have been designed and validated in the last two decades, which have served as tools to explore the above conundrums on how the evolution, size, diversity, and composition of a repertoire impact the selection of more specific and higher affinity antibodies to any given target. A first generation of man-made antibodies included all the antibody genes encoding the repertoire of circulating antibodies, but a relatively limited subset of the genes predominated the panning of the repertoires, showing that many genes are dispensable. A second generation of synthetic repertoires followed. Learning the lesson from the study of natural

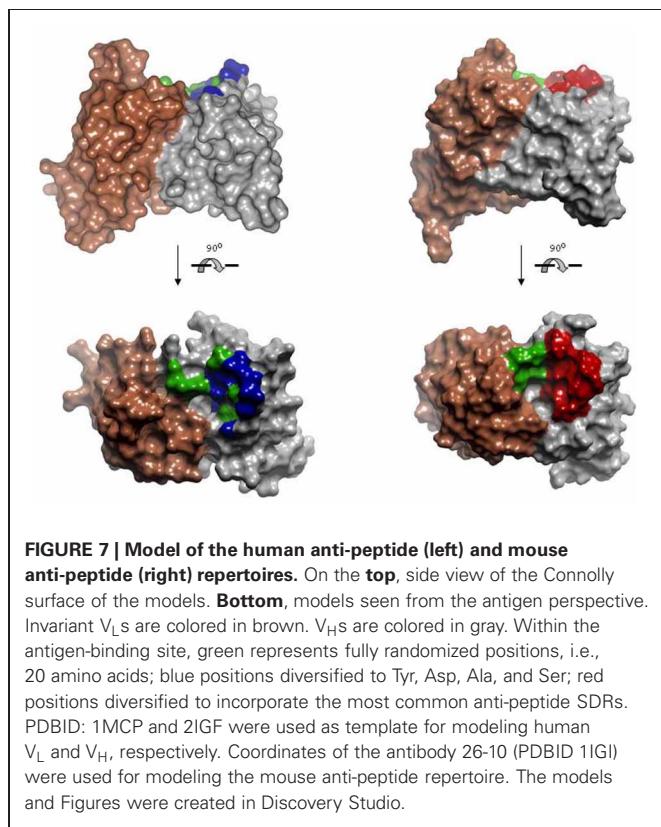


FIGURE 7 | Model of the human anti-peptide (left) and mouse anti-peptide (right) repertoires. On the **top**, side view of the Connolly surface of the models. **Bottom**, models seen from the antigen perspective. Invariant V_Ls are colored in brown. V_Hs are colored in gray. Within the antigen-binding site, green represents fully randomized positions, i.e., 20 amino acids; blue positions diversified to Tyr, Asp, Ala, and Ser; red positions diversified to incorporate the most common anti-peptide SDRs. PDBID: 1MCB and 2IGF were used as template for modeling human V_L and V_H, respectively. Coordinates of the antibody 26–10 (PDBID 1IGI) were used for modeling the mouse anti-peptide repertoire. The models and Figures were created in Discovery Studio.

and man-made naïve repertoires, the new generation of synthetic human antibody repertoires was built on single or a few well-behaved scaffolds. The diversity in these libraries was designed to mimic that of the natural antibodies. These repertoires produced antibodies to a vast array of the diverse antigens. In a further round of design and testing, minimalist repertoires have been designed and validated. These repertoires have been designed to display antigen-binding sites made of very few amino acids or even only binary Y/S and W/S mixes, again yielding antibodies specific against diverse antigens. Finally, rational repertoires encoding genes with predefined recognition features have been tested, which hold the promise of increasing the probability of obtaining more specific and higher affinity antibodies.

Looking forward into the future, new technologies such as next generation sequencing (NGS) are providing the means

to study whole natural (Weinstein et al., 2009; Jiang et al., 2011) and man-made repertoires (Glanville et al., 2009) in expedited ways and at relatively low costs (Fischer, 2011; Benichou et al., 2012). Having access to the complete information encoded in repertoires before and after selection under diverse selection pressures, combined with faster and more accurate 3D modeling methods (Almagro et al., 2011; Kuroda et al., 2012) and indeed new conceptual tools and algorithms such as network analysis, may reveal new features of antibody repertoires. These findings will hopefully further impact the theories addressing the origin and evolution of antibody binding specificity. It will also inform the design and optimization of man-made repertoires to isolate more potent, stable, safe and efficacious antibody-based therapeutics, at a lower cost.

REFERENCES

- Abhinandan, K. R., and Martin, A. C. (2010). Analysis and prediction of VH/VL packing in antibodies. *Protein Eng. Des. Sel.* 23, 689–697.
- Aitken, R., Gilchrist, J., and Sinclair, M. C. (1997). A single diversified VH gene family dominates the bovine immunoglobulin repertoire. *Biochem. Soc. Trans.* 25, 326S.
- Al-Lazikani, B., Lesk, A. M., and Chothia, C. (1997). Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.* 273, 927–948.
- Almagro, J. C. (2004). Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires. *J. Mol. Recognit.* 17, 132–143.
- Almagro, J. C., Beavers, M., Hernandez-Guzman, F., Maier, J., Shaulsky, J., Butenhof, K., et al. (2011). Antibody modeling assessment. *Proteins* 79, 3050–3066.
- Almagro, J. C., and Fransson, J. (2008). Humanization of antibodies. *Front. Biosci.* 13, 1619–1633.
- Almagro, J. C., Hernandez I., del Carmen Ramirez, M., and Vargas-Madrazo, E. (1997). The differences between the structural repertoires of IGHV germ-line gene segments of mice and humans: implication for the molecular mechanism of the immune response. *Mol. Immunol.* 34, 1199–1214.
- Almagro, J. C., Raghunathan, G., Beil, E., Janecki, D. J., Chen, Q., Dinh, T., et al. (2012). Characterization of a high-affinity human antibody with a disulfide bridge in the third complementarity-determining region of the heavy chain. *J. Mol. Recognit.* 25, 125–135.
- Alt, F., and Baltimore, D. (1982). Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions. *Proc. Natl. Acad. Sci. U.S.A.* 79, 4118–4122.
- Amzel, L. M., and Poljak, R. J. (1979). Three-dimensional structure of immunoglobulins. *Annu. Rev. Biochem.* 48, 961–997.
- Barbas, C. F. 3rd., Amberg, W., Simoncits, A., Jones, T. M., and Lerner, R. A. (1993). Selection of human anti-hapten antibodies from semisynthetic libraries. *Gene* 137, 57–62.
- Barbas, C. F. 3rd., Bain, J. D., Hoekstra, D. M., and Lerner, R. A. (1992). Semisynthetic combinatorial antibody libraries: a chemical solution to the diversity problem. *Proc. Natl. Acad. Sci. U.S.A.* 89, 4457–4461.
- Barthelemy, P. A., Raab, H., Appleton, B. A., Bond, C. J., Wu, P., Wiesmann, C., et al. (2008). Comprehensive analysis of the factors contributing to the stability and solubility of autonomous human VH domains. *J. Biol. Chem.* 283, 3639–3654.
- Benichou, J., Ben-Hamo, R., Louzoun, Y., and Efroni, S. (2012). Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135, 183–191.
- Betz, A. G., Rada, C., Pannell, R., Milstein, C., and Neuberger, M. S. (1993). Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity, and specific hot spots. *Proc. Natl. Acad. Sci. U.S.A.* 90, 2385–2388.
- Birtalan, S., Fisher, R. D., and Sidhu, S. S. (2010). The functional capacity of the natural amino acids for molecular recognition. *Mol. Biosyst.* 6, 1186–1194.
- Brezsinschek, H. P., Foster, S. J., Dorner, T., Brezsinschek, R. I., and Lipsky, P. E. (1998). Pairing of variable heavy and variable kappa chains in individual naive and memory B cells. *J. Immunol.* 160, 4762–4767.
- Carnemolla, B., Neri, D., Castellani, P., Leprini, A., Neri, G., Pini, A., et al. (1996). Phage antibodies with pan-species recognition of the oncofoetal angiogenesis marker fibronectin ED-B domain. *Int. J. Cancer* 68, 397–405.
- Chailyan, A., Marcatili, P., Cirillo, D., and Tramontano, A. (2011). Structural repertoire of immunoglobulin λ light chains. *Proteins* 79, 1513–1524.
- Chothia, C., and Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* 196, 901–917.
- Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., et al. (1992). Structural repertoire of the human VH segments. *J. Mol. Biol.* 227, 799–817.
- Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., et al. (1989). Conformations of immunoglobulin hypervariable regions. *Nature* 342, 877–883.
- Clark, L. A., Ganesan, S., Papp, S., and Vlijmen, H. W. V. (2006). Trends in antibody sequence changes during the somatic hypermutation process. *J. Immunol.* 177, 333–340.
- Cobaugh, C., Almagro, J., Pogson, M., Iverson, B., and Georgiou, G. (2008). Synthetic antibody libraries focused towards peptide ligands. *J. Mol. Biol.* 378, 622–633.
- Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221, 709–713.
- Cox, J. P., Tomlinson, I. M., and Winter, G. (1994). A directory of human germ-line V kappa segments

- reveals a strong bias in their usage. *Eur. J. Immunol.* 24, 827–836.
- Dal-Bo, M., Giudice, I. D., Bomben, R., Capello, D., Bertoni, F., Forconi, F., et al. (2011). B-cell receptor, clinical course and prognosis in chronic lymphocytic leukaemia: the growing saga of the IgHV3 subgroup gene usage. *Br. J. Haematol.* 153, 3–14.
- Davies, D. R., and Metzger, H. (1983). Structural basis of antibody function. *Annu. Rev. Immunol.* 1, 87–117.
- Davies, J., and Riechmann, L. (1994). ‘Camelising’ human antibody fragments: NMR studies on VH domains. *FEBS Lett.* 339, 285–290.
- De Genst, E., Saerens, D., Muyldermans, S., and Conrath, K. (2006). Antibody repertoire development in camelids. *Dev. Comp. Immunol.* 30, 187–198.
- De Genst, E., Silence, K., Decanniere, K., Conrath, K., Loris, R., Kinne, J., et al. (2006). Molecular basis for the preferential cleft recognition by dromedary heavy-chain antibodies. *Proc. Natl. Acad. Sci. U.S.A.* 103, 4586–4591.
- de Wildt, R. M., Hoet, R. M., van Venrooij, W. J., Tomlinson, I. M., and Winter, G. (1999). Analysis of heavy and light chain pairings indicates that receptor editing shapes the human antibody repertoire. *J. Mol. Biol.* 285, 895–901.
- Dooley, H., Flajnik, M. F., and Porter, A. J. (2003). Selection and characterization of naturally occurring single-domain (IgNAR) antibody fragments from immunized sharks by phage display. *Mol. Immunol.* 40, 25–33.
- Dörner, T., Foster, S. J., Farner, N. L., and Lipsky, P. E. (1998). Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. *Eur. J. Immunol.* 28, 3384–3396.
- Ewert, S., Cambillau, C., Conrath, K., and Pluckthun, A. (2002). Biophysical properties of camelid V (HH) domains compared to those of human V (H)3 domains. *Biochemistry* 41, 3628–3636.
- Ewert, S., Honegger, A., and Pluckthun, A. (2004). Stability improvement of antibodies for extracellular and intracellular applications: CDR grafting to stable frameworks and structure-based framework engineering. *Methods* 34, 184–199.
- Ewert, S., Huber, T., Honegger, A., and Pluckthun, A. (2003). Biophysical properties of human antibody variable domains. *J. Mol. Biol.* 325, 531–553.
- Farner, N. L., Dorner, T., and Lipsky, P. E. (1999). Molecular mechanisms and selection influence the generation of the human V lambda J lambda repertoire. *J. Immunol.* 162, 2137–2145.
- Fellouse, F. A., Barthelemy, P. A., Kelley, R. F., and Sidhu, S. S. (2006). Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code. *J. Mol. Biol.* 357, 100–114.
- Fellouse, F. A., Esaki, K., Birtalan, S., Raptis, D., Cancasci, V. J., Koide, A., et al. (2007). High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. *J. Mol. Biol.* 373, 924–940.
- Fellouse, F. A., Li, B., Compaan, D. M., Peden, A. A., Hymowitz, S. G., and Sidhu, S. S. (2005). Molecular recognition by a binary code. *J. Mol. Biol.* 348, 1153–1162.
- Fellouse, F. A., Wiesmann, C., and Sidhu, S. S. (2004). Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proc. Natl. Acad. Sci. U.S.A.* 101, 12467–12472.
- Fennell, B. J., Darmanin-Sheehan, A., Hufton, S. E., Calabro, V., Wu, L., Muller, M. R., et al. (2010). Dissection of the IgNAR V domain: molecular scanning and orthologue database mining define novel IgNAR hallmarks and affinity maturation mechanisms. *J. Mol. Biol.* 400, 155–170.
- Finlay, W. J., DeVore, N. C., Dobrovolskaia, E. N., Gam, A., Goodear, C. S., and Slater, J. E. (2005). Exploiting the avian immunoglobulin system to simplify the generation of recombinant antibodies to allergenic proteins. *Clin. Exp. Allergy* 35, 1040–1048.
- Finlay, W. J., Shaw, I., Reilly, J. P., and Kane, M. (2006). Generation of high-affinity chicken single-chain Fv antibody fragments for measurement of the Pseudomonas aeruginosa pungens toxin domoic acid. *Appl. Environ. Microbiol.* 72, 3343–3349.
- Fischer, N. (2011). Sequencing antibody repertoires: the next generation. *MAbs* 3, 17–20.
- Fisher, R. D., Ultsch, M., Lingel, A., Schaefer, G., Shao, L., Birtalan, S., et al. (2010). Structure of the complex between HER2 and an antibody paratope formed by side chains from tryptophan and serine. *J. Mol. Biol.* 402, 217–229.
- Flajnik, M. F., Deschacht, N., and Muyldermans, S. (2011). A case of convergence: why did a simple alternative to canonical antibodies arise in sharks and camels? *PLoS Biol.* 9:e1001120. doi: 10.1371/journal.pbio.1001120
- Flajnik, M. F., and Dooley, H. (2009). The generation and selection of single-domain, v region libraries from nurse sharks. *Methods Mol. Biol.* 562, 71–82.
- Foote, J., and Winter, G. (1992). Antibody framework residues affecting the conformation of the hypervariable loops. *J. Mol. Biol.* 224, 487–499.
- Gill, D. S., and Damle, N. K. (2006). Biopharmaceutical drug discovery using novel protein scaffolds. *Curr. Opin. Biotechnol.* 17, 653–658.
- Gilliland, G., Luo, J., Vafa, O., and Almagro, J. (2012). Leveraging SBDD in protein therapeutic development: antibody engineering. *Methods Mol. Biol.* 841, 321–349.
- Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G. R., et al. (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20216–20221.
- Govaert, J., Pellis, M., Deschacht, N., Vincke, C., Conrath, K., Muyldermans, S., et al. (2012). Dual beneficial effect of interloop disulfide bond for single domain antibody fragments. *J. Biol. Chem.* 287, 1970–1979.
- Griffiths, A. D., Malmqvist, M., Marks, J. D., Bye, J. M., Embleton, M. J., McCafferty, J., et al. (1993). Human anti-self antibodies with high specificity from phage display libraries. *EMBO J.* 12, 725–734.
- Griffiths, A. D., Williams, S. C., Hartley, O., Tomlinson, I. M., Waterhouse, P., Crosby, W. L., et al. (1994). Isolation of high affinity human antibodies directly from large synthetic repertoires. *EMBO J.* 13, 3245–3260.
- Hackel, B. J., Ackerman, M. E., Howland, S. W., and Wittrup, K. D. (2010). Stability and CDR composition biases enrich binder functionality landscapes. *J. Mol. Biol.* 401, 84–96.
- Hackel, B. J., and Wittrup, K. D. (2010). The full amino acid repertoire is superior to serine/tyrosine for selection of high affinity immunoglobulin G binders from the fibronectin scaffold. *Protein Eng. Des. Sel.* 23, 211–219.
- Hamers-Casterman, C., Atarhouch, T., Muyldermans, S., Robinson, G., Hamers, C., Songa, E. B., et al. (1993). Naturally occurring antibodies devoid of light chains. *Nature* 363, 446–448.
- Harding, F. A., Stickler, M. M., Razo, J., and Dubridge, R. B. (2010). The immunogenicity of humanized and fully human antibodies: residual immunogenicity resides in the CDR regions. *MAbs* 2, 256–265.
- Harmsen, M. M., and De Haard, H. J. (2007). Properties, production, and applications of camelid single-domain antibody fragments. *Appl. Microbiol. Biotechnol.* 77, 13–22.
- Harmsen, M. M., Ruuls, R. C., Nijman, I. J., Niewold, T. A., Frenken, L. G., and de Geus, B. (2000). Llama heavy-chain V regions consist of at least four distinct subfamilies revealing novel sequence features. *Mol. Immunol.* 37, 579–590.
- Hoogenboom, H. R. (2005). Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.* 23, 1105–1116.
- Huang, S. C., Jiang, R., Glas, A. M., and Milner, E. C. (1996). Non-stochastic utilization of Ig V region genes in unselected human peripheral B cells. *Mol. Immunol.* 33, 553–560.
- Ignatovich, O., Tomlinson, I. M., Jones, P. T., and Winter, G. (1997). The creation of diversity in the human immunoglobulin V (lambda) repertoire. *J. Mol. Biol.* 268, 69–77.
- Jarutat, T., Frisch, C., Nickels, C., Merz, H., and Knappik, A. (2006). Isolation and comparative characterization of Ki-67 equivalent antibodies from the HuCAL phage display library. *Biol. Chem.* 387, 995–1003.
- Jarutat, T., Nickels, C., Frisch, C., Stellmacher, F., Hofig, K. P., Knappik, A., et al. (2007). Selection of vimentin-specific antibodies from the HuCAL phage display library by subtractive panning on formalin-fixed, paraffin-embedded tissue. *Biol. Chem.* 388, 651–658.
- Jiang, N., Weinstein, J., Penland, L., White, R. R., Fisher, D., and Quake, S. R. (2011). Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc. Natl. Acad. Sci. U.S.A.* 108, 5348–5353.
- Johansson, J., Aveskogh, M., Munday, B., and Hellman, L. (2002). Heavy chain V region diversity in the duck-billed platypus (*Ornithorhynchus anatinus*): long and highly variable complementarity-determining region 3 compensates for limited germline diversity. *J. Immunol.* 168, 5155–5162.
- Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S., and Winter, G. (1986). Replacing the complementarity-determining regions in a human antibody with

- those from a mouse. *Nature* 321, 522–525.
- Kabat, E. A., and Wu, T. T. (1971). Attempts to locate complementarity-determining residues in the variable positions of light and heavy chains. *Ann. N.Y. Acad. Sci.* 190, 382–393.
- Knappik, A., Ge, L., Honegger, A., Pack, P., Fischer, M., Wellnhofer, G., et al. (2000). Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J. Mol. Biol.* 296, 57–86.
- Köhler, G., and Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* 256, 495–497.
- Koide, S., and Sidhu, S. S. (2009). The importance of being tyrosine: lessons in molecular recognition from minimalist synthetic binding proteins. *ACS Chem. Biol.* 4, 325–334.
- Kuroda, D., Shirai, H., Jacobson, M., and Nakamura, H. (2012). Computer-aided antibody design. *Protein Eng. Des. Sel.* 25, 507–522.
- Lee, C. V., Hymowitz, S. G., Wallweber, H. J., Gordon, N. C., Billeci, K. L., Tsai, S. P., et al. (2006). Synthetic anti-BR3 antibodies that mimic BAFF binding and target both human and murine B cells. *Blood* 108, 3103–3111.
- Lee, C. V., Liang, W. C., Dennis, M. S., Eigenbrot, C., Sidhu, S. S., and Fuh, G. (2004). High-affinity human antibodies from phage-displayed synthetic Fab libraries with a single framework scaffold. *J. Mol. Biol.* 340, 1073–1093.
- Lefranc, M. P., Giudicelli, V., Kaas, Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., et al. (2005). IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* 33, D593–D597.
- Lerner, R. (2011). Rare antibodies from combinatorial libraries suggests an S.O.S. component of the human immunological repertoire. *Mol. Biosyst.* 7, 1004–1012.
- Li, F., and Aitken, R. (2004). Cloning of porcine scFv antibodies by phage display and expression in *Escherichia coli*. *Vet. Immunol. Immunopathol.* 97, 39–51.
- Li, H., Jiang, Y., Prak, E. L., Radic, M., and Weigert, M. (2001). Editors and editing of anti-DNA receptors. *Immunity* 15, 947–957.
- Lo Conte, L. L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* 285, 2177–2198.
- MacCallum, R., Martin, A., and Thornton, J. (1996). Antibody-antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.* 262, 732–745.
- Maizels, N. (2005). Immunoglobulin gene diversification. *Annu. Rev. Genet.* 39, 23–46.
- Marget, M., Sharma, B. B., Tesar, M., Kretzschmar, T., Jenisch, S., Westphal, E., et al. (2000). Bypassing hybridoma technology: HLA-C reactive human single-chain antibody fragments (scFv) derived from a synthetic phage display library (HuCAL) and their potential to discriminate HLA class I specificities. *Tissue Antigens* 56, 1–9.
- Marks, J., Hoogenboom, H., Bonnert, T., McCafferty, J., Griffiths, A., and Winter, G. (1991). By-passing immunization. Human antibodies from V-gene libraries displayed on phage. *J. Mol. Biol.* 222, 581–597.
- Martin, A. C., and Thornton, J. M. (1996). Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J. Mol. Biol.* 263, 800–815.
- Martinez-Jean, C., Folch, G., and Lefranc, M. (2001). Nomenclature and overview of the mouse (*Mus musculus* and *Mus* sp.) immunoglobulin kappa (IGK) genes. *Exp. Clin. Immunogenet.* 18, 255–279.
- Matsuda, F., Ishii, K., Bourvagnet, P., Kuma, K., Hayashida, H., Miyata, T., and Honjo, T. (1998). The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J. Exp. Med.* 188, 2151–2162.
- McCafferty, J., Griffiths, A. D., Winter, G., and Chiswell, D. J. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* 348, 552–554.
- Mian, I., Bradwell, A., and Olson, A. (1991). Structure, function and properties of antibody binding sites. *J. Mol. Biol.* 217, 133–151.
- Morea, V., Tramontano, A., Rustici, M., Chothia, C., and Lesk, A. M. (1998). Conformations of the third hyper-variable region in the VH domain of immunoglobulins. *J. Mol. Biol.* 275, 269–294.
- Morrison, S. L., Johnson, M. J., Herzenberg, L. A., and Oi, V. T. (1984). Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains. *Proc. Natl. Acad. Sci. U.S.A.* 81, 6851–6855.
- Muyldermans, S., Baral, T. N., Retamozzo, V. C., De Baetselier, P., De Gens, E., Kinne, J., et al. (2009). Camelid immunoglobulins and nanobody technology. *Vet. Immunol. Immunopathol.* 128, 178–183.
- Muyldermans, S., and Lauwerys, M. (1999). Unique single-domain antigen binding fragments derived from naturally occurring camel heavy-chain antibodies. *J. Mol. Recognit.* 12, 131–140.
- Nelson, A., Dhimolea, E., and Reichert, J. (2010). Development trends for human monoclonal antibody therapeutics. *Nat. Rev. Drug Discov.* 9, 767–774.
- Neri, D., Petrucci, H., and Roncucci, G. (1995). Engineering recombinant antibodies for immunotherapy. *Cell Biophys.* 27, 47–61.
- Neri, D., Pini, A., and Nissim, A. (1998). Antibodies from phage display libraries as immunochemical reagents. *Methods Mol. Biol.* 80, 475–500.
- Neuberger, M. S., and Milstein, C. (1995). Somatic hypermutation. *Curr. Opin. Immunol.* 7, 248–254.
- Nguyen, V. K., Desmyter, A., and Muyldermans, S. (2001). Functional heavy-chain antibodies in Camelidae. *Adv. Immunol.* 79, 261–296.
- Nguyen, V. K., Hamers, R., Wyns, L., and Muyldermans, S. (1999). Loss of splice consensus signal is responsible for the removal of the entire C (H)1 domain of the functional camel IGG2A heavy-chain antibodies. *Mol. Immunol.* 36, 515–524.
- Nguyen, V. K., Su, C., Muyldermans, S., and van der Loo, W. (2002). Heavy-chain antibodies in Camelidae; a case of evolutionary innovation. *Immunogenetics* 54, 39–47.
- Nishibori, N., Horiuchi, H., Furusawa, S., and Matsuda, H. (2006). Humanization of chicken monoclonal antibody using phage-display system. *Mol. Immunol.* 43, 634–642.
- Noia, J. D., and Neuberger, M. (2007). Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* 76, 1–22.
- North, B., Lehmann, A., and Dunbrack, R. J. (2011). A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* 406, 228–256.
- O'Brien, P. M., Aitken, R., O'Neil, B. W., and Campo, M. S. (1999). Generation of native bovine mAbs by phage display. *Proc. Natl. Acad. Sci. U.S.A.* 96, 640–645.
- Ohara, R., Knappik, A., Shimada, K., Frisch, C., Ylera, F., and Koga, H. (2006). Antibodies for proteomic research: comparison of traditional immunization with recombinant antibody technology. *Proteomics* 6, 2638–2646.
- Padlan, E. A. (1977). Structural basis for the specificity of antibody-antigen reactions and structural mechanisms for the diversification of antigen-binding specificities. *Q. Rev. Biophys.* 10, 35–65.
- Padlan, E. A. (1994). Anatomy of the antibody molecule. *Mol. Immunol.* 31, 169–217.
- Parvari, R., Avivi, A., Lentner, F., Ziv, E., Tel-Or, S., Burstein, Y., et al. (1988). Chicken immunoglobulin gamma-heavy chains: limited VH gene repertoire, combinatorial diversification by D gene segments and evolution of the heavy chain locus. *EMBO J.* 7, 739–744.
- Parvari, R., Ziv, E., Lantner, F., Heller, D., and Schechter, I. (1990). Somatic diversification of chicken immunoglobulin light chains by point mutations. *Proc. Natl. Acad. Sci. U.S.A.* 87, 3072–3076.
- Parvari, R., Ziv, E., Lantner, F., Tel-Or, S., Burstein, Y., and Schechter, I. (1987a). A few germline genes encode the variable regions of chicken immunoglobulin light and gamma-heavy chains. *Prog. Clin. Biol. Res.* 238, 15–26.
- Parvari, R., Ziv, E., Lentner, F., Tel-Or, S., Burstein, Y., and Schechter, I. (1987b). Analyses of chicken immunoglobulin light chain cDNA clones indicate a few germline V lambda genes and allotypes of the C lambda locus. *EMBO J.* 6, 97–102.
- Persson, H., Lantto, J., and Ohlin, M. (2006). A focused antibody library for improved hapten recognition. *J. Mol. Biol.* 357, 607–620.
- Pini, A., Viti, F., Santucci, A., Carnemolla, B., Zardi, L., Neri, P., et al. (1998). Design and use of a phage display library. Human antibodies with subnanomolar affinity against a marker of angiogenesis eluted from a two-dimensional gel. *J. Biol. Chem.* 273, 21769–21776.
- Prassler, J., Steidl, S., and Urlinger, S. (2009). *In vitro* affinity maturation of HuCAL antibodies: complementarity determining region exchange and RapMAT technology. *Immunotherapy* 1, 571–583.
- Prassler, J., Thiel, S., Pracht, C., Polzer, A., Peters, S., Bauer, M., et al. (2011). HuCAL PLATINUM, a synthetic Fab library optimized for sequence diversity and superior performance in mammalian expression systems. *J. Mol. Biol.* 413, 261–278.
- Raaphorst, F. M., Raman, C. S., Nall, B. T., and Teale, J. M. (1997). Molecular mechanisms

- governing reading frame choice of immunoglobulin diversity genes. *Immunol. Today* 18, 37–43.
- Rada, C., Ehrenstein, M. R., Neuberger, M. S., and Milstein, C. (1998). Hot spot focusing of somatic hypermutation in MSH2-deficient mice suggests two stages of mutational targeting. *Immunity* 9, 135–141.
- Rader, C., Ritter, G., Nathan, S., Elia, M., Gout, I., Jungbluth, A. A., et al. (2000). The rabbit antibody repertoire as a novel source for the generation of therapeutic human antibodies. *J. Biol. Chem.* 275, 13668–13676.
- Ragunathan, G., Smart, J., Williams, J., and Almagro, J. (2012). Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. *J. Mol. Recognit.* 25, 103–113.
- Rajewsky, K., Förster, I., and Cumano, A. (1987). Evolutionary and somatic selection of the antibody repertoire in the mouse. *Science* 238, 1088–1094.
- Ramirez-Benitez, M. C., and Almagro, J. C. (2001). Analysis of antibodies of known structure suggests a lack of correspondence between the residues in contact with the antigen and those modified by somatic hypermutation. *Proteins* 45, 199–206.
- Ratcliffe, M. J. (2006). Antibodies, immunoglobulin genes and the bursa of Fabricius in chicken B cell development. *Dev. Comp. Immunol.* 30, 101–118.
- Reichert, J. (2012). Marketed therapeutic antibodies compendium. *MAbs* 4, 413–415.
- Reynaud, C. A., Anquez, V., Dahan, A., and Weill, J. C. (1985). A single rearrangement event generates most of the chicken immunoglobulin light chain diversity. *Cell* 40, 283–291.
- Reynaud, C. A., Anquez, V., Grimal, H., and Weill, J. C. (1987). A hyperconversion mechanism generates the chicken light chain preimmune repertoire. *Cell* 48, 379–388.
- Reynaud, C. A., Anquez, V., and Weill, J. C. (1991). The chicken D locus and its contribution to the immunoglobulin heavy chain repertoire. *Eur. J. Immunol.* 21, 2661–2670.
- Reynaud, C. A., Dahan, A., Anquez, V., and Weill, J. C. (1989). Somatic hyperconversion diversifies the single V κ gene of the chicken with a high incidence in the D region. *Cell* 59, 171–183.
- Reynaud, C. A., Dahan, A., and Weill, J. C. (1983). Complete sequence of a chicken lambda light chain immunoglobulin derived from the nucleotide sequence of its mRNA. *Proc. Natl. Acad. Sci. U.S.A.* 80, 4099–4103.
- Riechmann, L., and Muyldermans, S. (1999). Single domain antibodies: comparison of camel VH and camelized human VH domains. *J. Immunol. Methods* 231, 25–38.
- Rothe, C., Urlinger, S., Lohning, C., Prassler, J., Stark, Y., Jager, U., et al. (2008). The human combinatorial antibody library HuCAL GOLD combines diversification of all six CDRs according to the natural immune system with a novel display method for efficient selection of high-affinity antibodies. *J. Mol. Biol.* 376, 1182–1200.
- Sanz, I. (1991). Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions. *J. Immunol.* 147, 1720–1729.
- Saphire, E., Parren, P., Pantophlet, R., Zwick, M., Morris, G., Rudd, P., et al. (2001). Crystal structure of a neutralizing human IgG against HIV-1: a template for vaccine design. *Science* 293, 1155–1159.
- Schäble, K., Thiebe, R., Bensch, A., Brensing-Küppers, J., Heim, V., Kirschbaum, T., et al. (1999). Characteristics of the immunoglobulin V κ genes, pseudogenes, relics and orphans in the mouse genome. *Eur. J. Immunol.* 29, 2082–2086.
- Schäble, K., Thiebe, R., Flügel, A., Meindl, A., and Zachau, H. (1994). The human immunoglobulin kappa locus: pseudogenes, unique and repetitive sequences. *Biol. Chem. Hoppe Seyler* 375, 189–199.
- Schroeder, H. W. Jr. (2006). Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev. Comp. Immunol.* 30, 119–135.
- Schroeder, H. W. Jr., Hillson, J. L., and Perlmutter, R. M. (1987). Early restriction of the human antibody repertoire. *Science* 238, 791–793.
- Schroeder, H. W. Jr., Hillson, J. L., and Perlmutter, R. M. (1990). Structure and evolution of mammalian VH families. *Int. Immunol.* 2, 41–50.
- Schroeder, H. W. Jr., Ippolito, G. C., and Shiokawa, S. (1998). Regulation of the antibody repertoire through control of HCDR3 diversity. *Vaccine* 16, 1383–1390.
- Schroeder, H. W. Jr., Mortari, F., Shiokawa, S., Kirkham, P. M., Elgavish, R. A., and Bertrand, F. E. 3rd. (1995). Developmental regulation of the human antibody repertoire. *Ann. N.Y. Acad. Sci.* 764, 242–260.
- Shi, L., Wheeler, J. C., Sweet, R. W., Lu, J., Luo, J., Tornetta, M., et al. (2010). *In vitro* selection of high-affinity antibodies from synthetic fab libraries displayed on phage as pIX fusion proteins. *J. Mol. Biol.* 397, 385–396.
- Shirai, H., Kidera, A., and Nakamura, H. (1996). Structural classification of CDR-H3 in antibodies. *FEBS Lett.* 399, 1–8.
- Sidhu, S. S., and Fellouse, F. A. (2006). Synthetic therapeutic antibodies. *Nat. Chem. Biol.* 2, 682–688.
- Sidhu, S. S., Li, B., Chen, Y., Fellouse, F. A., Eigenbrot, C., and Fuh, G. (2004). Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions. *J. Mol. Biol.* 338, 299–310.
- Silacci, M., Brack, S. S., Spath, N., Buck, A., Hillinger, S., Arni, S., et al. (2006). Human monoclonal antibodies to domain C of tenascin-C selectively target solid tumors *in vivo*. *Protein Eng. Des. Sel.* 19, 471–478.
- Silacci, M., Brack, S., Schirru, G., Marlind, J., Ettorre, A., Merlo, A., et al. (2005). Design, construction, and characterization of a large synthetic human antibody phage display library. *Proteomics* 5, 2340–2350.
- Sinclair, M. C., Gilchrist, J., and Aitken, R. (1997). Bovine IgG repertoire is dominated by a single diversified VH gene family. *J. Immunol.* 159, 3883–3889.
- Smith, G. P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* 228, 1315–1317.
- Stanfield, R. L., Dooley, H., Flajnik, M. F., and Wilson, I. A. (2004). Crystal structure of a shark single-domain antibody V region in complex with lysozyme. *Science* 305, 1770–1773.
- Stanfield, R. L., Dooley, H., Verdino, P., Flajnik, M. F., and Wilson, I. A. (2007). Maturation of shark single-domain (IgNAR) antibodies: evidence for induced-fit binding. *J. Mol. Biol.* 367, 358–372.
- Stanfield, R. L., and Wilson, I. A. (2010). “Antibody molecular structure,” in *Therapeutic Monoclonal Antibodies: From Bench to Clinic*, ed Z. An (John Wiley and Sons, Inc.), 51–66.
- Stephens, S., Emtage, S., Vetterlein, O., Chaplin, L., Bebbington, C., Nesbitt, A., et al. (1995). Comprehensive pharmacokinetics of a humanized antibody and analysis of residual anti-idiotypic responses. *Immunology* 85, 668–674.
- Thiebe, R., Schäble, K., Bensch, A., Brensing-Küppers, J., Heim, V., Kirschbaum, T., et al. (1999). The variable genes and gene families of the mouse immunoglobulin kappa locus. *Eur. J. Immunol.* 29, 2072–2081.
- Tomlinson, I. M., Cox, J. P., Herardi, G. E., Lesk, A. M., and Chothia, C. (1995). The structural repertoire of the human V kappa domain. *EMBO J.* 14, 4628–4638.
- Tomlinson, I. M., Walter, G., Jones, P. T., Dear, P. H., Sonnhammer, E. L., Winter, G. (1996). The imprint of somatic hypermutation on the repertoire of human germline V genes. *J. Mol. Biol.* 256, 813–817.
- Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B., and Winter, G. (1992). The repertoire of human germline VH sequences reveals about fifty groups of VH segments with different hypervariable loops. *J. Mol. Biol.* 227, 776–798.
- Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature* 14, 575–581.
- Tsurushita, N., Park, M., Pakabunto, K., Ong, K., Avdalovic, A., Fu, H., et al. (2004). Humanization of a chicken anti-IL-12 monoclonal antibody. *J. Immunol. Methods* 295, 9–19.
- Vandenbroucke, K., de Haard, H., Beirnaert, E., Dreier, T., Lauwerys, M., Huyck, L., et al. (2010). Orally administered L. lactis secreting an anti-TNF Nanobody demonstrate efficacy in chronic colitis. *Mucosal Immunol.* 3, 49–56.
- Vargas-Madrazo, E., Lara-Ochoa, F., and Almagro, J. C. (1995). Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions associated to the mechanism of immune recognition. *J. Mol. Biol.* 254, 497–504.
- Virnekas, B., Ge, L., Pluckthun, A., Schneider, K. C., Wellnhofer, G., and Moroney, S. E. (1994). Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucleic Acids Res.* 22, 5600–5607.
- Viti, F., Nilsson, F., Demartis, S., Huber, A., and Neri, D. (2000). Design and use of phage display libraries for the selection of antibodies and enzymes. *Meth. Enzymol.* 326, 480–505.
- Weill, J. C., and Reynaud, C. A. (1992). Early B-cell development in chickens, sheep and rabbits. *Curr. Opin. Immunol.* 4, 177–180.

- Weinstein, J., Jiang, N., White, R. R., Fisher, D., and Quake, S. (2009). High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324, 807–810.
- Wesolowski, J., Alzogaray, V., Reyelt, J., Unger, M., Juarez, K., Urrutia, M., et al. (2009). Single domain antibodies: promising experimental and therapeutic tools in infection and immunity. *Med. Microbiol. Immunol.* 198, 157–174.
- Williams, A., and Barclay, A. (1988). The immunoglobulin superfamily-domains for cell surface recognition. *Annu. Rev. Immunol.* 6, 381–405.
- Wilson, I. A., and Stanfield, R. L. (1994). Antibody-antigen interactions: new structures and new conformational changes. *Curr. Opin. Struct. Biol.* 4, 857–867.
- Wilson, P. C., Bouteiller, O. D., Liu, Y.-J., Potter, K., Banchereau, J., Capra, J. D., et al. (1998). Somatic hypermutation introduces insertions and deletions into immunoglobulin V genes. *J. Exp. Med.* 187, 59–70.
- Wong, S. E., Sellers, B. D., and Jacobson, M. P. (2011). Effects of somatic mutations on CDR loop flexibility during affinity maturation. *Proteins* 79, 821–829.
- Wu, L., Oficjalska, K., Lambert, M., Fennell, B. J., Darmanin-Sheehan, A., Ni Shuilleabhain, D., et al. (2011). Fundamental characteristics of the immunoglobulin VH repertoire of chickens in comparison with those of humans, mice, and camelids. *J. Immunol.* 188, 322–333.
- Wu, T., Johnson, G., and Kabat, E. (1993). Length distribution of CDRH3 in antibodies. *Proteins* 16, 1–7.
- Wu, T. T., and Kabat, E. A. (1970). An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* 132, 211–250.
- Yamanaka, H. I., Inoue, T., and Ikeda-Tanaka, O. (1996). Chicken monoclonal antibody isolated by a phage display system. *J. Immunol.* 157, 1156–1162.
- Zemlin, M., Klinger, M., Link, J., Zemlin, C., Bauer, K., Engler, J. A., et al. (2003). Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J. Mol. Biol.* 334, 733–749.
- Zhai, W., Glanville, J., Fuhrmann, M., Mei, L., Ni, I., Sundar, P. D., et al. (2011). Synthetic antibodies designed on natural sequence landscapes. *J. Mol. Biol.* 412, 55–71.
- Zhao, S., and Lu, J. (2010). A germline knowledge based computational approach for determining antibody complementarity determining regions. *Mol. Immunol.* 47, 694–700.
- Zhu, D., Lossos, C., Chapman-Fredricks, J., Matthews, J., Ikkatt, O., Ruiz, P., et al. (2011). Biased use of the IGHV4 family and evidence for antigen selection in *Chlamydophila psittaci*-negative ocular adnexal extranodal marginal zone lymphomas. *PLoS ONE* 6:e29114. doi: 10.1371/journal.pone.0029114

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 August 2012; paper pending published: 05 September 2012; accepted: 27 October 2012; published online: 15 November 2012.

Citation: Finlay WJJ and Almagro JC (2012) Natural and man-made V-gene repertoires for antibody discovery. *Front. Immun.* 3:342. doi: 10.3389/fimmu.2012.00342

This article was submitted to Frontiers in B Cell Biology, a specialty of Frontiers in Immunology.

Copyright © 2012 Finlay and Almagro. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing

Miri Michaeli¹, Hila Noga¹, Hilla Tabibian-Keissar^{1,2}, Iris Barshack² and Ramit Mehr^{1*}

¹ The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel

² Department of Pathology, Sheba Medical Center and Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel

Edited by:

Harry W. Schroeder, University of Alabama at Birmingham, USA

Reviewed by:

Harry W. Schroeder, University of Alabama at Birmingham, USA

Michael Zemlin, Philipps University Marburg, Germany

*Correspondence:

Ramit Mehr, The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan 52900, Israel.
e-mail: ramit.mehr@biu.ac.il

High-throughput sequencing (HTS) yields tens of thousands to millions of sequences that require a large amount of pre-processing work to clean various artifacts. Such cleaning cannot be performed manually. Existing programs are not suitable for immunoglobulin (Ig) genes, which are variable and often highly mutated. This paper describes Ig High-Throughput Sequencing Cleaner (Ig-HTS-Cleaner), a program containing a simple cleaning procedure that successfully deals with pre-processing of Ig sequences derived from HTS, and Ig Insertion—Deletion Identifier (Ig-Indel-Identifier), a program for identifying legitimate and artifact insertions and/or deletions (indels). Our programs were designed for analyzing Ig gene sequences obtained by 454 sequencing, but they are applicable to all types of sequences and sequencing platforms. Ig-HTS-Cleaner and Ig-Indel-Identifier have been implemented in Java and saved as executable JAR files, supported on Linux and MS Windows. No special requirements are needed in order to run the programs, except for correctly constructing the input files as explained in the text. The programs' performance has been tested and validated on real and simulated data sets.

Keywords: B cell receptor, computer programs, high-throughput sequencing, immunoglobulin (Ig) genes, insertions and deletions (indels)

INTRODUCTION

Studying the generation, development and selection of lymphocyte repertoires, and their functions during immune responses, is essential for understanding the function of the immune system in healthy individuals, and in monitoring and intervening with the immune system in immune deficient, autoimmune disease or cancer patients. The recent development of high-throughput sequencing (HTS) enables researchers to obtain large numbers of sequences from several samples simultaneously (Galan et al., 2010). HTS has a great advantage over classical sequencing methods in the field of immunoglobulin (Ig) gene research, as it enables us to extract more sequences per sample and it is sensitive enough so we can identify different unique sequences. Ig genes encode the B cell receptors (BCR) and tend to accumulate point mutations in their sequences in order to improve the BCRs affinity to antigens. Mutation analysis, which is one aspect of Ig gene research, enables the tracking of mutation accumulation in the BCRs and hence analysis of the diversification of Ig gene sequences that originate from the same ancestor. Thus, HTS presents us now, for the first time, with the ability to analyze and compare large samples of mutated Ig gene repertoires in health, aging and disease (Campbell et al., 2008; Boyd et al., 2009; Gibson et al., 2009; Scheid et al., 2009; Ademokun et al., 2010; Dunn-Walters and Ademokun, 2010). However, the huge numbers of sequences obtained require a large amount of pre-processing work to clean out artifacts, sort sequences according to sample according to their molecular

identification (MID) tags, identify primers, and discard sequences that do not contain enough information, such as sequences much shorter or longer than the expected length of an Ig variable region gene, or sequences with average quality scores below a defined threshold. Several programs are already used by the scientific community to study the B cell mutational patterns, such as SoDA (Volpe et al., 2006), SoDA2 (Munshaw and Kepler, 2010), and iHMmune-Align (Gaëta et al., 2007) which perform V(D)J segment identification; identification of clonally-related Ig gene sequences using clustering methods (Chen et al., 2010); ClustalW2 (Larkin et al., 2007), for alignment of clonally-related sequences; and IgTree (Barak et al., 2008), for creating lineage trees from the sets of aligned clonally-related sequences. However, in order to use these programs and receive reliable results that are not affected by sequencing artifacts, one must first make sure that all such artifacts are cleaned out of the input data.

Although HTS has already been available for several years, there are very few such cleaning programs available for users, and none that can deal with the cleaning of Ig genes. For example, the program CANGS (Pandey et al., 2010) has a very good pipeline of cleaning sequences, but it discards unique sequences and searches for primers and MID tags with perfect matches only, while it is known that tags and primers are often incomplete or sequenced incorrectly. Another program that could be used for Ig gene data cleaning is SeqTrim (Falgueras et al., 2010). This program does all the desired cleaning processes, but has two

main disadvantages: one is that it runs on sequences that were inserted in vectors, so the program identifies the inserts only according to vector sequences found in databases such as NCBI's UniVec, EMBL's emvec, or BLAST. The user does not have the option to insert the ends of the genes, e.g., primers, as an input. Therefore, this program is suitable only for researchers that use sequencing with vectors. A second disadvantage is that SeqTrim requires other external programs. Ngs_backbone (Blanca et al., 2011) is another program that can be used for cleaning Ig gene sequences, but it requires external softwares. Additionally, there are several programs that trim the adapters (primers used in HTS) and the template-specific primers (MID tags or barcodes). One of them is TagCleaner (Schmieder et al., 2010), but this program takes all the input sequences, aligns them in order to identify the most frequent sequences at both ends that are supposed to be the adapters/tags and trims them. TagCleaner and similar programs are therefore ineffective when sequences come from several samples and hence contain several tag combinations; tags are composed of different sequences, so no consensus sequence can be deduced correctly using this method. In addition, highly homologous or, in contrast, highly variable sequences may yield erroneous alignments and therefore false identification of adapters. We have tested TagCleaner on our Ig genes data but this program could not correctly identify the tags. Another program is TagDust (Lassmann et al., 2009), but this program identifies artifact reads by comparing all reads to a library of sequences and checking for significant match. It is not possible to create such library for Ig gene sequences, as they undergo somatic hypermutation (SHM) in a high rate, and thus can diverge (Cook and Tomlinson, 1995; Rajewsky, 1996). EA-utils (Aronesty, 2011), Scythe (Buffalo, n.d.), SeqPrep (John, n.d.), FASTX (Gordon, n.d.), and Trim Galore! (Krueger, n.d.) are additional programs that are used to trim adapters among other functions; however, it is not possible to identify adapters in the 5' end of the reads and to search for multiple different adapters using these programs. Using trimLRPatterns [one tool of ShortRead, (Morgan et al., 2009)] lacks the option to search for multiple different adapters. Trimmomatic (Bolger and Giorgi, n.d.) does not allow identifying adapters in the 5' end of the reads, and anyway is compatible to Illumina sequencing only. FAR [The Flexible Adapter Remover, (Unknown, n.d.)] is capable of searching for multiple adapters using a simple global alignment algorithm, but it does not record the combination of adapters (or barcodes or MIDs) if found and cut. This is important when sequencing several different samples in the same sequencing run. Cutadapt (Martin, 2011) offers an easy-to-use command-line program that searches for multiple adapters and trim them, and is specialized for small RNA sequences. However, cutadapt currently does not support using a configuration file in which, for example, a list of adapters can be specified; hence, inputting several adapter sequences is via the command-line, which makes it slightly cumbersome. AdapterRemoval (Lindgreen, 2012) can search for multiple adapters in both 5' and 3' ends of the reads and discards reads that do not exceed a minimum length given by the user. However, none of the above mentioned programs assign the reads to their original samples according to their MIDs (barcodes), although the search of adapters should be similar to the

identification of MIDs. Btrim (Kong, 2011) presents the closest cleaning options to our desired ones. In addition to trimming adapters and low quality regions as some of the above programs do, it can also identify barcodes and assign the reads to their original samples. However, Btrim has several shortcomings. First, it is limited to Linux. Second, similar to some of the above-mentioned programs, it requires some knowledge regarding the use with the command-line. A program with a user interface or even a double-click program is preferable, as its use can be included in an automated pipeline easy to operate even by users with little experience with computers. Moreover, it can search for multiple adapters or barcodes, and it can work with a configuration file containing all the adapters or barcodes to search, but it requires this file to contain pairs of 5' and 3' adapters or barcodes. This way, if barcodes were used in several combinations for samples, as we do, this file should contain all possible combinations of barcodes.

Thus, we needed—and created—a program that can clean the sequences of artifacts, and would be suitable for use with Ig genes in spite of their unique characteristics. We present here the Ig-HTS-Cleaner program, which enables the user to give the ends of the genes (primers and MID tags) as input no matter what their origin is, can handle multiple tags, and does not require any additional programs in order to run. Our Ig-HTS-Cleaner program does not require any knowledge in programming nor complicated installation, only a simple input file which contains the parameters for run. Moreover, the FASTA output files enable easy downstream analyses of the sequences.

Sequencing of complete Ig genes can currently be carried out only by the 454 pyrosequencing sequencing platform or the illumina platform. The reason is the maximum sequencing length required in order to get the full Ig gene. Only 454 or illumina currently reach a maximum read length of 500 nucleotides. Other platforms can reach such lengths only by using the paired-ends method, when only the ends of the gene are sequenced and the middle is inferred by comparing to a reference gene. This, of course, is not valid with Ig genes, due to their huge variability and the lack of a reference gene. Other platforms require assembly of complete sequences from shorter reads, which is also a problem due to the high mutation load and large numbers of similar but not identical sequences in Ig genes. When other sequencing platforms reach the same read length, our programs may be used on the data generated by them as well.

One of the shortcomings of pyrosequencing is that during the sequencing of homopolymer tracts (HPTs, repeats of the same nucleotide), the polymerase can add or delete one or more nucleotides from these repeats, or alternatively, the signal of poly-nucleotide incorporation can be misread (Huse et al., 2007). These errors may result in insertions/deletions (indels) that are a result of the sequencing and therefore are considered as artifacts (Margulies et al., 2005). In Ig gene research, it is very important to distinguish between artifact indels and legitimate indels that are a result of normal SHM and affinity maturation of B cells, although naturally occurring indels are very rare. Legitimate indels should be taken into account when analyzing mutations of the B cell Ig genes, and artifact indels

should be discarded from the analysis. There are several common approaches for dealing with indels. Campbell et al. used an algorithm that discards sequences with insertions, deletions or substitutions that occurred near or in HPTs, unless the indel or the substitution was seen in both the forward and reverse reads (Campbell et al., 2008). In other studies, all sequences with any type of indel are excluded from analysis (Boyd et al., 2010; Wu et al., 2010) or included without accounting for indels (Wu et al., 2010). CANGS (Pandey et al., 2010) identifies indels that appear only in HPTs near the primers and discards them. The program does not identify indels occurring far from the primers. VarScan (Koboldt et al., 2009) and VARiD (Dalca et al., 2010) can identify indels, but these programs deal with variability and single nucleotide polymorphism (SNP) identification, and do not distinguish between legitimate and artifact indels, and are hence less suitable for identifying and discarding artifact indels from Ig genes. Recently, two methods for distinguishing true indels from sequencing artifacts have been developed. Dindel (Albers et al., 2011) utilizes a probabilistic method that accounts for the increased indel rates near HPTs. However, Dindel detects indels from short reads generated by Illumina sequencing and aligns the reads to a specific region in the genome. Hence, Dindel is not suitable for Ig genes because they are longer than Illumina reads and they cannot be aligned to a specific region in the genome due to the enormous variability and randomness of Ig gene rearrangements. PiCALL (Bansal and Libiger, 2011) also detects indels using a probabilistic method, but it works on a population of diploid individuals. Therefore, piCALL is not suitable for Ig gene sequences, in which different cells have different Ig genes.

Sometimes, discarding all the indels causes loss of information. For example, if an indel appears in a unique sequence, most of the algorithms would discard this sequence because no other sequence contains this indel. Nevertheless, there is a chance that this indel is legitimate, and we refer to such cases as uncertain indels. Because in several of our studies we perform sequencing on DNA from preserved tissue samples which do not yield large enough amounts of DNA (Tabibian-Keissar et al., 2008), we do not want to simply discard all unique sequences. Since there is no absolute way to identify all sequencing errors, and because we have limited sequencing data, we intend to save as many sequences as possible, and still avoid as many artificial indels and point mutations as possible. To address this issue, we developed Ig-Indel-Identifier, a program that identifies legitimate and artifact indels and does not discard sequences that contain uncertain indels. Hence, we may decide to keep these uncertain sequences for maximal information utilization.

This paper describes both Ig-HTS-Cleaner and Ig-Indel-Identifier. Our programs were designed to process Ig genes, but they are easily applicable to all types of sequences.

MATERIALS AND METHODS

Ig-HTS-Cleaner and Ig-Indel-Identifier have been programmed in Java on the Windows operating system and is saved as an executable JAR file. The JAR files support Linux and Microsoft Windows. An executable file for each of our programs is available at <http://immsilico2.lnx.biu.ac.il/Software.html>. To execute

the program, the user should double-click on the program symbol after saving it in the same directory where the input files are saved. No special requirements are needed in order to run them, except for correctly constructing the input files as explained below.

Ig-HTS-Cleaner INPUT AND OUTPUT

The program receives as an input the following files, which should be saved in the same directory as the program. (1) A group of *.fna files (*denotes any desirable name) containing the FASTA sequence reads—see **Figure 1** for the structure of a typical read. (2) Quality (*.qual) files containing the scores for the sequences (both file types are received from the sequencing platform). (3) An input.txt file created and updated by the user, containing the parameters for the cleaning, such as the MID tags and the primer sequences, the length range within which the sequence is considered legitimate, the samples (in case more than one sample was sequenced), etc. A detailed explanation on how to fill the input.txt file can be found in **Figure 2**. The input file should be created precisely according to these instructions.

As output, the program generates the following files:

1. FailedInFindMIDs.txt—a FASTA file containing all the sequences in which the program could not find one or both MID tags, or the MID combination did not match the table in the input file.
2. FailedInFindPrimers.txt—a FASTA file containing all the sequences in which the program could not find one or both primers, such that only sequences with identifiable MIDs and primers are included in further analysis.
3. FailedInCheckLength.txt—a FASTA file containing all the sequences for which L2 (sequence length between the primers) was not within the allowed range.
4. FailedInQuality.txt—contains all sequences that had identifiable primers at both ends and were within the allowed length range, but the average quality score of the sequence was lower than the input threshold.
5. Log.txt—a tab-delimited text file that assembles details of the run in two parts. The first part (**Figure 3A**) contains a table of the samples that were sequenced and the following numbers per each sample: total number of sequences found, number of failed sequences in finding primers, percent of the last value out of the total, number of failed sequences in length, percent of the last value out of total, and out of the total after the previous stage, number of sequences with lower quality than threshold, percent of the last value out of total and out of the total after the previous stage, total remaining number of sequences, average quality score. The second part (**Figure 3B**) contains information regarding the total numbers of the run, such as (partial list): how many sequences failed in the MID tags finding step, how many failed in the primers finding step, how many failed in the length check, how many failed in the quality check, how many are in the sense or anti-sense orientation, and a short review of the parameters of the run.
6. A text file for each sample, containing the sequences that passed all the checks and were identified as belonging to that

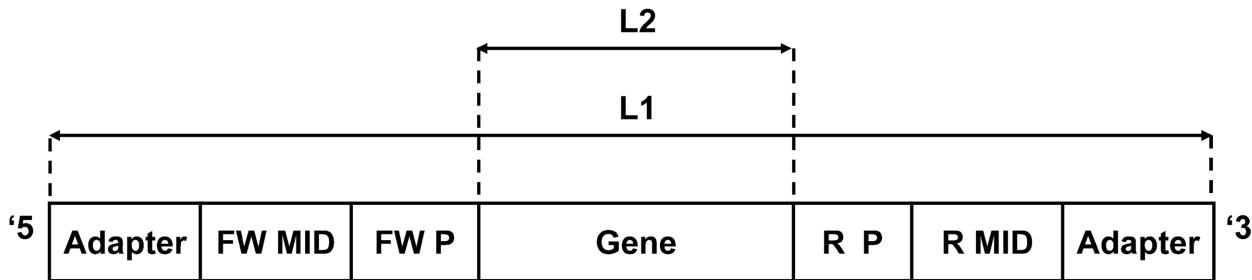


FIGURE 1 | A typical sequencer output (e.g., 454) read consists of the following segments (ordered from the 5' end to the 3' end): adapter: sequencer adapters. FW MID denotes the forward molecular identifier (tag). FW P denotes the forward polymerase chain reaction (PCR) primer. Gene is the target sequence. R P denotes the reverse PCR primer. R MID denotes the reverse molecular identifier (tag). L1: Sequence length before first filtering—should be in the allowed

range defined by the first minimum and maximum values ($\min L1 < L1 < \max L1$, for example, 200 and 400 in **Figure 2**). L2: Sequence length after primers were cut—should be in the allowed range defined by the second minimum and maximum values ($\min L2 < L2 < \max L2$, for example, 150 and 360 in **Figure 2**). These ranges can be changed in the input file according to the data set and the specific requirements of each study.

- sample according to their MID tag combination (given as input, **Figure 2**).
7. A text file for each sample, containing the quality score sequence of each sequence belonging to the sample, with the scores referring to the gene sequence only (without scores for the MID tags, primers, etc.).

Ig-HTS-Cleaner ALGORITHM

The program works according to the following outline (**Figure 4**):

1. Read input file and initiate parameters for the run.
2. Read .fna and .qual files and parse them for the run.
3. For each sequence, if it is longer or shorter than the allowed range given in the input file, discard it from analysis.
4. For each sequence that passed the previous step, do the following:
 - 4.1. Find tags (MIDs) at both ends of the sequence. If found, go to 4.2, else go to 4.5.
 - 4.2. Find Primers at both ends of the sequence. If found, go to 4.3, else go to 4.5.
 - 4.3. Check whether the length of the sequence between the primers is within the input range. If so, go to 4.4, else go to 4.5.
 - 4.4. Check the average quality score of the sequence. If the average quality score is below the input threshold, go to 4.5, else go to 4.6.
 - 4.5. A sequence that failed in one of the above stages will be written to a discard file according to the reason of failure.
 - 4.6. A sequence that succeeded in all the above stages will be written to an output file according to its MID combination.

MID tag finding

The goal of the first step is to find the MID tags for each sequence. Each 454 run enables the sequencing of several, pooled samples, as long as we mark each sequence according to its sample using known combinations of MID tags at both ends of the sequence

(because sequencing may start at each end of the sequence). To use this feature, the genes must undergo preliminary PCR with primers that are connected to specific 10-base oligonucleotides, representing the MID tags. These tags were composed by the sequencing company to provide different oligosequences that are distinguishable. In addition, the reads also contain adapter sequences at their ends that are required for starting the sequencing (**Figure 1**). Since the number of tags that are sufficiently distinguishable from each other (see below) is limited, one can use primers with a known combination of forward and reverse MID tags for each sample. This way, one can obtain different sequences from many different samples at one run, and later correctly attribute the sequences to the original samples by identifying the MID tag combination in each sequence. If either MID tag cannot be identified, the program would not be able to attribute the sequence to a sample and thus will reject it as a failed sequence. The program searches each sequence for a perfect match to each MID tag (taken from the input file) at both ends of the sequence, which are the regions where we expect the MID tags to be found (and not in the middle of the read, for example). The search is executed on a limited range of nucleotides, given in the input file. If a perfect match is achieved, the number of the tag is noted. If not, the program searches for a perfect match of the minimal MID length that the user has inserted in the input file. This minimal MID length represents the number of consecutive nucleotides of the tag on the side closer to the primers (the inner side of the sequence) that enables unambiguous identification of the tag. This is done because the sequencer more often inserts errors close to the sequence boundaries, so the tags might have been trimmed or contain errors at their outer edges. We found that five consecutive nucleotides are the lowest number of nucleotides that can still distinguish between the different MID tags we used (basic set—Hamming distance: 6, **Table 1**). However, the minimal MID length is dependent on the Hamming distance of the MID set used, and should be assigned correctly by the user. The program prioritizes a perfect match, thus in case of a perfect match to the minimal MID length of one MID tag, the program will continue the search

```

organism
human
chain
h
quality threshold
20
maximum mismatches allowed
2
fraction of primer to search
0.75
range to search primers in
50
minimal mid length
5
mids
ACGAGTGCCT
ACGCTCGACA
AGACGCACTC
#
forward
TGCAGMCAGGCCYGGACAAR
ARGRAAGGCCCTGGAGTGG
CCGCCAGGCTCAGGSAAG
MGGGAAGGGRCTGGAGTGG
GAAAGGCCTGGAGTGGATGGG
TTGAGTGGCTGGRAGGAC
#
reverse
TGACCRKGGTHCCYTGGCCC
#
minimum length
200
150
maximum length
400
360
table
1 1 sample1
2 2 sample2
3 3 sample3
1 2 sample4
2 3 sample5
1 3 sample6
3 2 sample7
2 1 sample8
3 1 sample9
#

```

FIGURE 2 | An example of the input.txt file content. In bold- words/characters that should always appear. **Organism**—represents the organism to which the sequences belong, “**Human**” in this example. **Chain**—represents the chain of the Ig (h, heavy as in this example; l, lambda; k, kappa). **Quality threshold**—the minimal average score for a sequence allowed. **Maximum mismatches allowed**—the number of mismatches the user allows when primers are being searched. “2” means that when primers are being searched, the sequence can contain 2 insertions/deletions or substitutions in the primer’s sequence. **Fraction of primer to search**—in case the full primer was not found, the program searches only the given fraction of the primer from the side closer to the gene. **Range to search primers in**—the search is executed on a limited range of bases at the ends of the read. **Minimal MID length**—in case the full MID was not found, the program searches for a perfect match of the minimal length of the MID from the side closer to the gene. **Mids**—a list of the MID tags that have been used in the current sequencing. The program automatically numbers the MID tags according to the insertion order. At the end of each list, a “#” should appear, see example. In case no MIDs

(Continued)

FIGURE 2 | Continued

were used, leave only the title and the “#”. **Forward**—a list of the forward primers that have been used in the current sequencing, used for identification of the primers. At the end of each list, a “#” should appear, see example. **Reverse**—a list of the reverse primers that have been used in the current sequencing, used for identification of the primers. At the end of each list, a “#” should appear, see example. **Minimum length**—two values, the first is the minimal length for the first filtering of the data (minL1). The second value represents the minimal length that is legitimate for the genes in between the primers (minL2). **Maximum length**—two values, the first is the maximal length for the first filtering of the data (maxL1). The second value represents the maximal length that is allowed for the genes in between the primers (maxL2). **Table**—contains the MID tag combination per each sample that was sequenced. MID tag numbers should coordinate with their serial number in the above list. This enables the program to attribute each sequence to its corresponding sample. Each line should contain: number of the forward MID tag/tab/number of the reverse MID tag/tab/sample id (see example). At the end of each list, a “#” should appear, see example. In case no MIDs were used, put 0 as the number of forward and reverse MIDs.

for a perfect match of the rest of the MID tags. Although rare, in case a perfect match of one MID tag and a perfect match of the minimal MID length of other MID tags are found in the same read-end, the program prefers the perfect match of the full MID tag.

A legitimate sequence contains a legitimate combination of tags (according to the input file) with the tag at the 5' end found in the sense orientation and at the one 3' end found in the anti-sense orientation. Each search for tags is performed using both the tags and their complementary sequences, in order to identify tags at both ends. If a match of two tags, one at each edge of the sequence, is found, the tag numbers are noted and the sequence in between the MID tags is passed on to the next stage of cleaning. Otherwise, the sequence is discarded from further analysis, and written to a file containing all the sequences that failed in this stage. It is important to note that Ig-HTS-Cleaner can also work in case no MIDs are listed (for example, when the sequencing was carried out on a single sample). The program will then search directly for primers, but it is important that the input file is written properly as detailed in **Figure 2**.

Primer identification

In this stage, the sequence between the MID tags (after MID tags have been removed) is searched for primers at both ends. Again, the search is executed on a limited range of nucleotides, given in the input file. The program runs using the primer lists given in the input file and searches for a perfect match of both forward and reverse primers in the current sequence. The program searches both the primer and its complementary sequence. If one or both primers were not found with a perfect match, the program searches for a partial match between the sequence and the primers, after the latter are trimmed (from the side furthest from the gene) leaving a fraction of the original primers’ length, given in the input file (e.g., 75% of the original length). Sometimes PCR and/or sequencing trim the primer ends. Searching only the primer fraction closer to the gene enables the program to identify even primers which contain errors or were trimmed at the ends distal to the gene. This is useful when

A											
Sample	Total	Failed in primers	% out of total	Failed in length	% out of total	% out of remaining	Failed in quality	% out of total	% out of remaining	Total remaining	Average score
Sample1	2097	22	1.05	0	0	0	1	0.048	0.048	2074	21
Sample2	1579	760	48.13	0	0	0	1	0.063	0.122	818	21
Sample3	2461	73	2.97	0	0	0	0	0.000	0.000	2388	21
Sample4	664	240	36.14	0	0	0	0	0.000	0.000	424	23
Sample5	749	30	4.01	0	0	0	0	0	0	719	22

B											
Total number of reads: 641574 How many reads were not within the first length boundaries and were discarded from further analysis: 253030 How many did not have MIDs at both ends of the sequence: 3970 How many did not have primers at both ends of the sequence (either forward or reverse primer, or both): 4380 How many were not within the second length range: 258 How many were shorter than the range: 258 How many were longer than the range: 0 How many sequences had average quality scores below the threshold: 7 How many sequences are in a sense orientation out of the ok sequences: 67234 How many sequences are in a antisense orientation out of the ok sequences: 59665 How many succeeded in partial match of forward primer: 17519 How many succeeded in partial match of reverse primer: 18098 How many failed in finding a forward primer: 1280 How many failed in finding a reverse primer: 1856 How many had primers, but both were sense or antisense, which probably means these are chimeric sequences: 1549 How many have different MID combinations than those given in the input, out of those that have MIDs: 1572 How many out of the total sequences could not be identified by their MIDs (percent): 4 MIDs were searched 5 nts from the side closer to the primers in order to address cases of deletion of the edges. The partial match was based on the following criteria: taking 0.75 of primer length from the side closer to the gene, searching for it in the 50 nts at the edges of the gene- depending on the orientation of the primer, allowing a maximum of 2 mismatches (insertions/ deletions/ substitutions). The threshold for the average quality score per sequence was: 20											

FIGURE 3 | A sample of an Ig-HTS-Cleaner log.txt output file. (A) The first part of the file (after importing the .txt file into a table). “Sample” is the sample name, inserted in the input file. “Total” represents the total number of sequences received from the sequencer before cleaning (as counted after MID identification). “Failed in primers” represents the number of sequences without primers at both ends. “% out of total” means the % of the “Failed in primers” column out of the total column. “Failed in length” represents the number of sequences with length not in between the input range. “% out of total” means the % of the “Failed in length” column out of the total column. “% of the remaining” for the “Failed in length” column represents the % after we subtract the number of sequences failed in primers from the total and calculate the % from that. “Failed in quality” represents the number of

sequences with an average quality score below the threshold. “% out of total” means the % of the “Failed in quality” column out of the total column. “% of the remaining” for the “Failed in quality” column represents the % after we subtract the number of sequences failed in primers and in length from the total and calculate the % from that. “Total remaining” represents the number of sequences that have passed all cleaning steps successfully. “Average score” is the average score for the sample. First, the average score for each sequence is calculated by an average of the scores per base, given in the .qual files in the 454 output. Then, the average score of the sample is calculated as the average of the scores of all sequences belonging to the sample. **(B)** The second part of the file, containing information and statistics regarding the run.

the reads are from one sample and thus there was no use in MID tags. The partial match allows the number of mismatches per primer defined by the user input. The larger the number of allowed mismatches, the more erroneous primer identifications would occur. Therefore, one should decide on the maximal number of mismatches allowed for the data, according to this trade-off. We examined this parameter on data sets from human and mouse tissues (data not shown). We ran each data set in Ig-HTS-Cleaner with different values between 1 and 10 for the number of allowed mismatches. For our human data set, we found that a value of two allowed mismatches is the best cut-off

which minimizes both the loss of sequences and the gain of erroneous sequences with falsely identified primers. For our mouse data set, we found that a value of four allowed mismatches is the best cut-off. This type of analysis may be performed on other data sets with Ig-HTS-Cleaner, to decide on the best value for each data set.

The partial match uses dynamic programming of local alignment of the sequence with each forward or reverse primer, based on the Smith-Waterman algorithm. If both forward and reverse primers are found, the gene orientation is known, hence the MID tag order is known as well, so the sample identity is known.

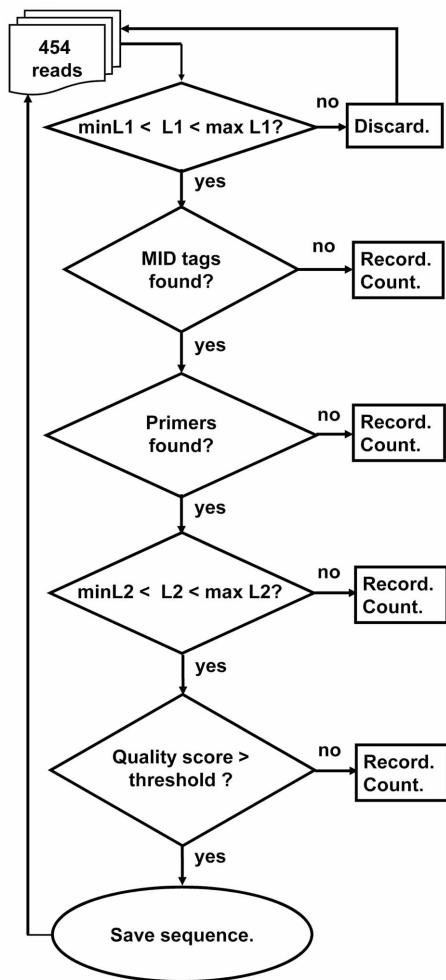


FIGURE 4 | A schematic outline of the Ig-HTS-Cleaner program algorithm. “Discard” means: discard the read because it is not likely to be an immunoglobulin gene sequence. “Record” means: write the sequence to a file containing all sequences that failed in this stage. “Count” means: count the number of failed sequences. In case the failure is in the last two stages, the counts are done per sample. “MID tags found?” means: “Is there an identifiable MID tag at each end of the sequence?” “Primers found?” means: “Is there an identifiable primer at each end of the sequence?” Checks of L1 and L2 are as described in Figures 1 and 2. “Quality score > threshold?” means: “Is the average quality score of the sequence larger than the threshold score given in the input file?” “Save sequence” means: “Write the sequence to the corresponding sample file. Advance the counter of sequences per sample.”

The sequence then proceeds to the next stage. If one or both primers were not found, the sequence is discarded from further analysis, and written to a file containing all the sequences that failed in this stage.

Length check

This is the last check before the sequence is accepted as legitimate. If the sequence length between the primers is within the allowed range, the sequence is written to the file of the sample corresponding to its MID tag combination; each file represents

Table 1 | List of MID tags, forward and reverse primers used for Ig-HTS-Cleaner validation^a.

	MID tags (5' to 3') ^b	ACGAGTGCCT ACGCTCGACA AGACGCACTC AGCACTGTAG ATCAGACACG ATATCGCGAG CGTGTCCTA CTCCGGTGTG TAGTATCAGC
Human	Forward primers (5' to 3')	TGCGMCAGGCCCYGGACAAR ARGRAGGCCCTGGAGTGG CCGCCAGGCTCCAGGSAAG MGGAAGGGRCTGGAGTGG GAAAGGCCTGGAGTGGATGGG TTGAGTGGCTGGGRAGGAC
	Reverse primer (5' to 3')	TGACCRKGTHCCYTGGCCC
Mouse—heavy chain	Forward primers (5' to 3')	AGRKYCARCTGCARCAAGYC TGCAGCTKMAKSAGTCAG GARGTGAAGCTKSTSGAGTC GAGGAGTCTGGAGGAGGCTT CTGGGATATTGCAGCCCTCC AGGTGTGCATTGTGAGGTGC GTSAGGTGCAAGCTKGTRGA CAATCCCAGGTTCACCTACAA
	Reverse primer (5' to 3')	GTGGTBCCSTS GCCCCAG
Mouse—light chain	Forward primers (5' to 3')	MTGATGACCCARTCTCCA SRGATATTGTGATGACGCAAGG AWTGTDCTSACCCARTCTCC CCTGTGGRGACATTGTGAT AYCCVGATGACYRTYCARATG CCAGATGTGAYRTYCARATG BCAGTGTGACATCCRVT ACACAGGCTCAGCTTCTCT TCCCAGGCTGTTGTGACTC CAACTGTGCTCACTCAGTC CTCTAGGAAGCACAGTCAAAC
	Reverse primer (5' to 3')	GTGGTBCCSTS GCCCCAG

^aKey to degenerate nucleotides: R = A + G; M = A + C; W = A + T; K = G + T; S = G + C; Y = C + T; H = A + T + C; B = G + T + C; D = G + A + T; N = A + C + G + T; V = G + A + C.

^bWe used the basic set, with Hamming distance = 6.

one sample. Otherwise, the sequence is discarded, and written to a file containing all the sequences that failed in this stage.

Quality check

For each sequence, a file containing the sequencing quality scores per each base is generated during the sequencing run. When using the 454 platform, each nucleotide in each sequence gets a score between 0 and 40 that represents the confidence level

of the sequencer that a specific nucleotide is the correct one. In other words, the higher the score per base (or average score per sequence), the better the quality of the sequencing. For each read, a sequence of numbers between 0 and 40 in the original length of the read is generated. The file with all the score sequences is the .qual file.

In addition to cleaning the sequence of tags and primers, a calculation of the quality score per sequence is performed in Ig-HTS-Cleaner using the .qual files. One can either look at each position to investigate its quality (for example, when examining point mutations), or can look at the average quality score per sequence or even per sample to evaluate the quality of sample sequencing (for example, when dealing with preserved tissue samples where DNA is often denatured). For each sequence, the quality score is calculated as the average score for all bases of the sequence, after tags and primers were removed. Then, the quality score of each sample is calculated, as the average score of the sample's sequences.

The calculation of minimal and average quality scores of one sample or a group of samples is also important for downstream analyses. For example, in Ig-Indel-Identifier (see below for more details), point mutations are checked for their quality score, and if the latter is lower than the threshold given by the user, the sequence can be discarded. In **Figure 5**, we show as an example the quality score analyses of four samples from human lymph nodes (LN).

In this example, many (42.81%) of the sequences in the LN samples had an average quality score of 37–38, which is considered a very high score (**Figure 5A**). On the other hand, many of the sequences also contained nucleotides with low quality scores (11–12, **Figure 5B**). These analyses led us to conclude that in our LN data, we could use the average quality score threshold of 30 without losing too many sequences. However, as most of the sequences appear to have a low minimal quality score (mostly lower than 20, with a peak in 11–12), we could not use a minimal threshold larger than 10.

Ig-Indel-Identifier INPUT AND OUTPUT

To identify indels created by the sequencer, the sequences are usually compared to some reference gene. In the case of Ig genes, where no reference gene can be used, the sequences should be organized by clones, according to their germline (GL) segment identifications. Then, a consensus sequence can be created for each clone based on all sequences in the clone, and serve as the reference gene. In our analysis pipeline, we first identify the GL segments for every sequence using SoDA (Volpe et al., 2006). Then we group the sequences (from the same sample) that use the same GL segments into one clone, and find the consensus sequence for the N-regions of this clone. In each position of the N-regions, the consensus GL contained the most frequent nucleotide in all the aligned sequences that belonged to the same clone. In the data presented below, we identified clones only based on their V(D)J GL segments for the purpose of demonstration of the action of Ig-Indel-Identifier. However, for proper data analysis, these groups of sequences are aligned and examined, as more than one clone may have the same V(D)J combination. For this purpose, one may use the clustering-based program

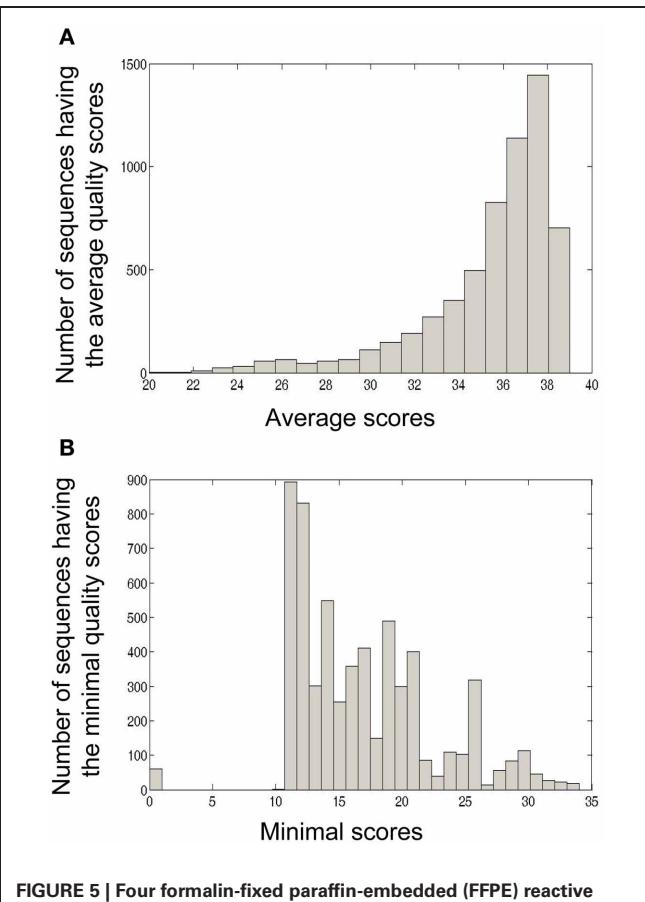


FIGURE 5 | Four formalin-fixed paraffin-embedded (FFPE) reactive lymph node tissues, archived about 10 years ago, from adult patients, were used as part of a study that dealt with Ig repertoire analysis of gastric lymphomas. In the current paper we used these samples in order to demonstrate quality scores obtained by the 454 platform using FFPE tissues. **(A)** Average quality score distribution of sequences obtained from four LN samples. For each sequence, the average quality score was calculated as the sum of all base scores divided by the length of the sequence. **(B)** Minimal quality score distribution of sequences obtained from four LN samples. For each sequence, the minimal quality score was calculated as the minimal base score value of all base scores in the sequence.

created by Gaeta et al. (Chen et al., 2010), which deals with groups of sequences that share the same V(D)J segments, and checks whether they belong to one or more clones. After the clonally-related groups of sequences are identified, the sequences from each clone are aligned, along with the “root” sequence composed of the GL segments and N-region consensus, using ClustalW2 (Larkin et al., 2007). The output files from ClustalW2 (*.txt files in the PIR format, which is an alignment format) serve as the input files for Ig-Indel-Identifier.

In addition to cleaning artifact indels, Ig-Indel-Identifier identifies point mutations (mismatches) by comparing to the reference gene or GL, and decides whether they are sequencing artifacts or may be derived from natural mutation processes. For each point mutation, Ig-Indel-Identifier checks the quality score of the base, given in the *.qual files. One of the parameters that are given by the user is the minimal quality score. This parameter

is used when Ig-Indel-Identifier compares the quality score of the point mutation to the given minimal quality score. If the quality score of the point mutation is lower than the user's threshold, and if this point mutation appears in fewer sequences in the clone than the number given in the input.txt file, the sequence is discarded. Moreover, Ig-Indel-Identifier identifies whether a point mutation occurred inside activation-induced cytidine deaminase (AID) motifs that contain HPT [AACAA or the complementary TGTT (MacCarthy et al., 2009)]. If so, the sequence shall not be discarded even if its quality score is lower than the threshold. If the user is not interested in identifying such mismatches, the value of the minimal quality score given in the input.txt file should be set to -1.

Ig-Indel-Identifier receives as input *.txt files, each containing an alignment of a group of clonally-related sequences with their consensus GL sequence; an *.input file, containing all the sequences from the current sample in the FASTA format; and a *.qual file, corresponding to the *.input file, which contains the quality score sequences of the sequences in *.input file. Ig-HTS-Cleaner generates these files automatically. In addition, the user should prepare a file called "input.txt," with integer values for three parameters, as follows:

1. The minimum number of sequences in a clone that must share the same indel or a low quality score point mutation for this indel or mutation to be considered legitimate. Before discarding a sequence, the user may decide that in case more sequences contain the same indel or low quality score point mutation, the suspected sequence shall not be discarded. The user can choose how many sequences in a clone must share that particular indel or mutation in order to save this sequence. The higher this threshold number, the fewer suspected sequences would be saved.
2. HPT length: the user can decide on the minimum length of same-nucleotide stretch that will be considered a HPT. The longer the HPT length, the more sequences would be saved, since fewer indels would be identified as near-HPT indels.
3. The minimal quality score for a point mutation to be considered legitimate. The higher the minimal quality score, the more sequences would be discarded. If the user is not interested in identifying such mismatches, the value of the minimal quality score given in the input.txt file should be set to -1.

To use Ig-Indel-Identifier on non-Ig gene sequences, one should create a consensus sequence of all sequences that should be checked. This can be done using several programs, such as ClustalW2 (Larkin et al., 2007). A consensus sequence is composed of the most frequent base in each position of the aligned sequences. Then, the consensus and the sequences should be run in ClustalW2 to create the *.txt file that contains the alignment. Ig-Indel-Identifier will work on this file together with the *.input file that should contain the sequences. When using ClustalW2 in order to align each of the sequences with its reference gene, each gap (of one or more nucleotides), in either the GL or the tested sequence, anywhere in the sequence, is designated as an indel that would be checked by Ig-Indel-Identifier.

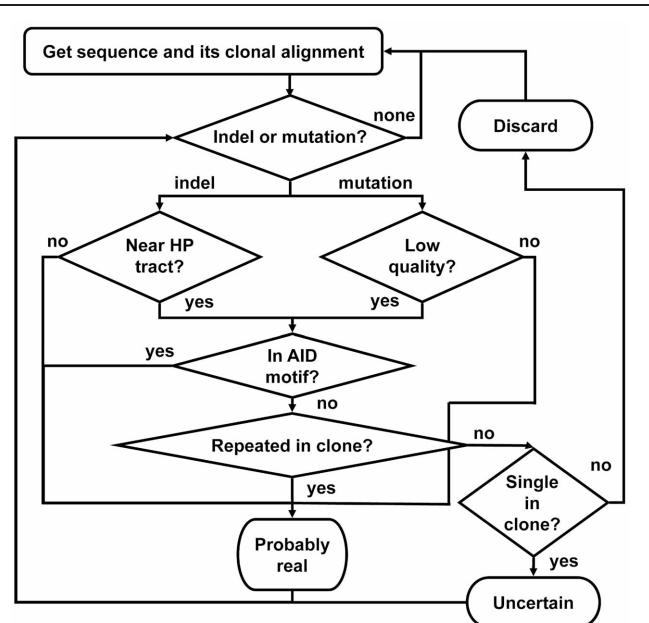
As output, the program generates the following files for each input file:

1. "(Name-of-input-file)-WithoutIndels.txt"—a FASTA file that contains a list of all sequences from the current input file which contain neither artifact nor uncertain indels.
2. "(Name-of-input-file)-CloneOfSize1WithIndels.txt"—a FASTA file that contains a list of sequences that contain suspected indels, but do not belong to a larger clone, so we cannot decide whether each indel is an artifact or not. We call these "uncertain indels," and keep these sequences separately, so one can perform the analysis either with or without them, as desired.
3. "(Name-of-input-file)-IllegitimateIndels.txt"—a FASTA file that contains a list of the sequences with artifact indels, which are part of a large clone, hence these indels are certainly artifact and not uncertain indels (this list has no overlap with output file number 2).
4. "(Name-of-input-file)-SeqsWithLowQualPointMuts.txt"—a FASTA file that contains a list of sequences that contain point mutations with quality scores lower than the minimal quality score given by the user, and which appeared in fewer sequences in the clone than set by the user.
5. "(Name-of-input-file)-Ig-Indel-Identifier.log"—a file containing the report of the run.

Ig-Indel-Identifier ALGORITHM

The program works according to the following outline (**Figure 6**):

1. For each sequence in each clone, do:
 - 1.1. Search for indels or point mutations by comparing the sequence to the corresponding GL (or consensus) sequence position by position, from the last indel or point mutation checked (or, in a new sequence—from the first position of the alignment) until the alignment ends. If an indel is found, go to 1.2. If a point mutation is found, go to 1.9, else go to 1.7.
 - 1.2. If this indel appears near a HPT (see below), go to 1.3, else go to 1.6.
 - 1.3. This indel is suspected to be an illegitimate (artifact) indel. If this indel is unique, i.e., no other sequence in its clone shares the same indel, go to 1.4, else go to 1.6.
 - 1.4. If the sequence is the single one in its clone go to 1.8, else go to 1.5.
 - 1.5. The sequence has an artifact indel and hence is discarded from further analyses. Write it to the appropriate file and go to the next sequence (step 1).
 - 1.6. This indel is considered as a legitimate indel. Go to 1.1.
 - 1.7. Sequence is OK. Write it to the appropriate file and go to the next sequence (step 1).
 - 1.8. Sequence has an uncertain indel. Mark as uncertain. Go to 1.1.
 - 1.9. Check if the quality score of the point mutation base is lower than the threshold given by the user as input. If so, go to 1.10, else go to 1.1.

**FIGURE 6 | A schematic outline of the Ig-Indel-Identifier program.**

"Indel or mutation?" means: "Are there (more) indels or mutations in the sequence?" "Low quality?" means: "Did the point mutation have a quality score lower than the input threshold?" "Near HP tract?" means: "Is the indel near a HPT?" "In AID motif?" means: "Is the indel or point mutation detected inside the known AID motif (AACCA or the complementary TGTT)?" "Repeated in clone?" means: "Are there more sequences in the clone with the same indel?" "Single in clone?" means: "Is this the only sequence in the clone?" "Discard" means: "Discard the sequence because it has an artifact indel." "Uncertain" means: save the sequence separately as it is suspected to have artifact indel(s), but there are no other sequences in this clone, hence we cannot be sure. "Probably real" means: save the sequence as it has a legitimate indel according to our definitions.

- 1.10. Check if the point mutation is inside the AID motif (AACCA or the complementary TGTT). If so, go to 1.11, else go to 1.11.
- 1.11. Check the number of sequences in the clone that share the same point mutation. If there are more sequences in the clone but no other sequence shares the same point mutation, go to 1.12. Else, go to 1.1.
- 1.12. The sequence has an artifact point mutation and hence is discarded from further analyses. Write it to the appropriate file and go to the next sequence (step 1).
2. Write the sequences to the output files.

In order to decide whether an indel is near or inside a HPT of a length that equals or exceeds the minimum length given by the user, we test the GL sequence to find whether one or more of the following conditions are fulfilled:

- The indel is inside a HPT.
- The indel is 5' to a HPT.
- The indel is 3' to a HPT.

We chose to first define HPTs by at least two identical nucleotides. Although this definition is very broad, as pairs of identical nucleotides are very common in all sequences, we preferred to check more indels than to leave sequences with illegitimate ones in the dataset. However, the user can choose the minimal length defining a HPT. If any of the above three conditions is fulfilled, this indel is suspected to be a sequencing artifact. A suspected indel is denoted as an artifact if fewer than the threshold number of sequences that share this indel exist in the clone.

Creating a simulated dataset for testing Ig-Indel-Identifier

In order to check the Ig-Indel-Identifier program, we collected 504 real sequences without indels from previous 454 HTS studies. This dataset contained either groups of clonally-related sequences or single sequences, all already aligned to their GL. We simulated the artificial induction of deletions (see below) on this dataset, creating a new and larger dataset of sequences, each with one deletion at most, near or inside HPTs of different lengths. The simulation was not created in order to reflect the "natural" generation of indels by the sequencer, but simply in order to have sequences with no more than one deletion, near or inside HPTs of different lengths, in order to test the Ig-Indel-Identifier program and analyze the results more easily. The simulation works as follows. For each sequence, the simulation decides whether to introduce a deletion or not. In case a deletion should be introduced, the simulation draws a value (2–10) for the length of HPT that the deletion should occur in. If the sequence does not contain any HPTs with that length, the simulation keeps drawing a value until at least one HPT with the drawn length is found in the sequence. From the list of positions contained in the chosen HPT, the simulation draws one position in one HPT to introduce the deletion in. A deletion (and indels in general) can occur 5' to the HPT, 3' to the HPT, or in its middle. Hence, the simulation draws the exact position for the deletion, according to the length of HPT drawn in the first step. Deletions were not allowed at the beginning or at the end of a sequence. Deletions were introduced by replacing the character (A/C/G/T) in the desired position of the aligned sequence by "-". Ig-Indel-Identifier identifies insertions and deletions according to the alignment of the clonally-related sequences and their GL, and hence a "-" sign in the GL is considered as an insertion, and in the sequence, it is considered as a deletion). Introducing only deletions during the simulation should not affect identification of indels by Ig-Indel-Identifier, and we used only deletions for convenience. After introducing a deletion to the sequence, the simulation draws a value (0–10) for the number of duplicate sequences carrying the same deletion that will be generated. One of our assumptions for correctly identifying artifact indels is that indels that appear in more than one sequence in the same clone are probably real and thus are not designated as suspect. The exact number of sequences carrying the same indel, which are needed for accurately identifying an indel, can change between datasets and observations. It is important to note that all the deletions introduced into the sequences during the simulation were artifact indels, and hence were all expected to be identified by Ig-Indel-Identifier and evaluated based on their frequency within the clone.

The simulation algorithm is as follows:

1. For each sequence (not GL) in a clone do:
 - 1.1. Decide whether the sequence will undergo a deletion. If so, go to 1.2. Else go to the next sequence (step 1).
 - 1.2. Draw a value (2–10) for the length of HPTs to search for in the sequence. If no HPT in the drawn value is found, repeat step 1.2. Else, go to 1.3.
 - 1.3. Draw one HPT from the sequence (there can be several, but only one is chosen).
 - 1.4. Draw the specific position near or inside the HPT (this depends on the HPT's length).
 - 1.5. Replace the character in the chosen position (A/C/G/T) with a “–”.
 - 1.6. Draw the number of duplicate sequences carrying the same indel (0–10).
 - 1.7. Generate the new sequence(s) and write them into a new file.

RESULTS

Ig-HTS-Cleaner—PERFORMANCE AND VALIDATION

We tested the performance of Ig-HTS-Cleaner on real data from 454 HTS. Twenty-nine DNA samples (from a study that will be published elsewhere) were subject to Ig gene amplification by PCR and the products were sequenced on the Roche 454 FLX Titanium platform to yield a total of 44,617 reads. We ran Ig-HTS-Cleaner on this data set with the following parameters: average quality score threshold of 20, 2 allowed mismatches in the primer search, 75% of the primer's length to search, and a range of 50 bases at the ends of the read for the MIDs and primers search (denoted as “combination 1,” see **Table 2**). Out of the 44,617 reads, 35,453 reads contained MID tags at both ends of the read. In the next step, Ig-HTS-Cleaner discarded 2504 sequences that did not contain identifiable primers at both ends, because in such sequences we cannot identify sequence orientation. It is important to identify primers at both ends, not only in order to identify where the gene is positioned inside the read, but also to identify the orientation of both primers, in order to discard those chimeric sequences—created during the PCR or the sequencing—that contain both primers in the same orientation rather than opposite orientations. These artifact sequences can be identified by primers with the same orientation. Only one read did not have a length within the requested range and was

discarded. The reason the latter number was so low is that all Ig genes are of similar lengths, such that if the whole gene between the primers was sequenced, it is highly likely to have the correct length. Much shorter or longer reads could be chimeric sequences (discussed below). Only seven sequences did not pass the average quality threshold, which we set to be 20. This is not surprising, as most of the nucleotides were sequenced with a quality score of 20–40 (see **Figure 5**). The user may decide whether to assign this threshold a higher value, and thus to discard more sequences. Finally, when Ig-HTS-Cleaner had finished running, we were left with 32,941 remaining sequences (**Table 2**, parameter combination 1). The list of MID tags and primers used in this specific run can be found in **Table 1**.

We ran Ig-HTS-Cleaner on the same data set with four different combinations of parameters (numbered 2–5 in **Table 2**), in order to demonstrate the influence of each parameter on the cleaning process. Parameter combination number 2 had the same parameter values as the original run (number 1) except for allowing up to 4 mismatches. It is not surprising that fewer reads were discarded in the stage of primer search (because more mismatches were allowed), thus there were more sequences attributed to MID tag combinations. In addition, six reads were included due to the lenient primer search, but two were discarded due to insufficient length, and 4 were discarded due to a low average quality score. Parameter combination number 3 had the same parameter values as the original run (number 1) except for the fraction of primer to search, which was set to 100%—that is, the program would search only for the full primer. It was obvious that in this case, more reads would be discarded, as we searched for the full primer sequence and allowed only 2 mismatches. There were fewer reads with low average quality scores, because some were already discarded in the primer search step. Parameter combination number 4 had the same parameter values as the original run (number 1) except for the range of 25 bases at the ends of the read for the MID and primers search. In this case, which took longer than previous runs (see below regarding run times), more reads were subjected to partial match, which allows mismatches, and thus fewer reads were discarded. Parameter combination number 5 had the same parameter values as the combination number 4 except for the fraction of primer to search, which was set to 100%. In this case, more reads were discarded than in the previous run, because the program searched only for the full primer. However, fewer reads were discarded than in runs 1 and 3, because more reads were subjected to partial match.

Table 2 | A summary of the Ig-HTS-Cleaner results in each parameter combination: human data set of 44K sequences.

Parameter combination	Number of sequences received	Number of sequences with tags	Number of sequences without primers	Number of sequences that failed due to length	Number of sequences that failed in quality check	Number of remaining sequences
1	44,617	35,453	2504	1	7	32,941
2	44,617	35,891	1662	3	11	34,215
3	44,617	35,114	3528	1	3	31,582
4	44,617	36,246	439	0	8	35,799
5	44,617	35,911	1684	0	4	34,223

See text for parameter combinations.

Validation of Ig-HTS-Cleaner results was done manually. Each cleaning step was examined individually. We checked for false negatives by looking at 50 reads that were discarded due to lack of MID tags and manually checked whether they do contain MID tags. None of the sequences was found to be false negative in this step. False positives were also not found when we performed manual checks on about 10 sequences from each sample (a total of ~200 sequences), which had successfully passed this step. The same validation steps were performed on ~150 sequences lacking primers and on ~100 sequences at both ends of which primers were found. This step was more complicated due to the use of dynamic programming for identifying primers with less than 100% match. About 50 sequences that did not contain 100% match of a primer were also checked to validate the dynamic programming algorithm's accuracy. All of the sequences that proceeded to the next cleaning step contained primers at both ends. Discarded sequences in this step indeed lacked one or more primers at their edges. Length checks were done automatically: a simple script validated that the lengths of all sequences that had passed the length check are truly within the allowed range, and that the sequences that had failed this step really were outside the allowed length range. We also validated that the sequences that were discarded due to lower quality score had indeed an average quality scores below the threshold. We collected these sequences and automatically calculated their average quality scores.

Applying Ig-HTS-Cleaner on the first data set of 44,617 reads, with a range of 50 nucleotides at each end to search primers in, took approximately 3–4 min to run on an Intel® core™2 CPU 6700, 2 GB RAM 2.67 GHz. When the primer search range was decreased to 25, Ig-HTS-Cleaner run took almost 1 h on the same computer. The longer run time is because when the primer search range decreases, primers that were located not in the first 25 bases but closer to the inner side of the sequence would not be found. Hence, the program would search for partial match of the primer(s), and that would take much longer. We then proceeded to test Ig-HTS-Cleaner on larger data sets, obtained from human and mouse DNA samples and together representing ~527,000 reads that were assigned into samples. However, for this number of reads we could not use the above-described computer due to memory shortage, and needed to run Ig-HTS-Cleaner on our UNIX server, which is equipped with larger RAM (16 GB). An Ig-HTS-Cleaner run on the ~527,000 reads took approximately 5 min on our UNIX server. Hence, we recommend using Ig-HTS-Cleaner on UNIX machines or on PCs with large internal memory. Regarding memory complexity, the program saves the reads for the whole running time in a special data structure,

representing $O(n \times k)$ memory, where n represents the length of a sequence, k represents the number of sequences, and $n \ll k$, hence $O(k)$ memory is required. For each instance of dynamic programming used in finding a partial primer match, we have $O(n \times m)$, where n represents the sequence length and $m \ll n$ represents the primer length. Actually, when the program searches for the primer, it searches it in a window shorter from the full sequence length at each side of the sequence, and not in the whole sequence, as we expect the primers to be on the sides of the sequence and not in the middle. Thus, $O(n \times m)$ is limited to a finite number. Moreover, a partial match search was carried out in less than 10% of the reads, reducing the complexity in one order of magnitude. To summarize, a run of ~500,000 sequences performed on a computer equipped with large internal memory (16 GB) would yield results within a short time (5 min). **Table 3** presents the cleaning results of both human and mouse data sets with an average quality score threshold of 20, and 2 and 4 allowed mismatches of primers for the human and mouse data sets, respectively. The two data sets were sequenced in the same run, but in different lanes. We present here the numbers of sequences attributed to each dataset and the cleaning results using Ig-HTS-Cleaner. The list of MID tags and primers used in this specific run can be found in **Table 1**.

Ig-Indel-Identifier—PERFORMANCE AND VALIDATION

We tested the performance of Ig-Indel-Identifier on the first dataset described above. The original study included 29 samples, but after cleaning with Ig-HTS-Cleaner, five samples out of the 29 yielded fewer than 30 sequences each, and these sequences were discarded from further analyses due to lack of interest. Data from 24 samples, which originally contained 36,944 sequences, were taken from the output of Ig-HTS-Cleaner. Out of these sequences, 33,767 sequences did not contain indels at all; this is reasonable, since SHM inserts mostly single base substitutions (Liu and Schatz, 2009; Steele, 2009). On the other hand, 3177 sequences contained indels (both uncertain and artifact), representing 8.6% of all sequences. Of the latter, 93 sequences with uncertain indels and 3084 with artifact indels were found (**Table 4**).

Applying Ig-Indel-Identifier on the data set of 44,617 reads took approximately 5 min to run on an Intel® core™2 CPU 6700, 2 GB RAM 2.67 GHz. Regarding memory complexity, the program saves the reads for the whole running time in a special data structure, representing $O(n \times k)$ memory, where n represents the length of a sequence and k represents the number of sequences. For each sequence, both the sequence and its GL (consensus) sequence are being compared, representing additional $O(n)$ memory space.

Table 3 | A summary of the Ig-HTS-Cleaner results: human and mouse data sets of 500 K sequences together.

Organism	Number of sequences with tags	Number of sequences without primers	Number of sequences that failed in due to length	Number of sequences that failed in quality check	Number of remaining sequences
Human	116,546	3248	4	10	113,284
Mouse	410,352	143,729	271	4	266,348

Table 4 | Numbers of sequences after Ig-Indel-Identifier cleaning.

Total	Number of sequences w/o indels	Total number of sequences with indels	% of sequences with indels	Number of uncertain indels ^a	Number of sequences with artifact indels ^b
36,944	33,767	3177	8.6	93	3084

^aAn uncertain indel is an indel in a single sequence that does not belong to a multi-sequence clone.

^bAn artifact indel is an indel near a HPT, where no other sequences in the same clone contain the same indel (and is not in a single sequence).

TESTING Ig-Indel-Identifier ON SIMULATED DATA

We collected 504 sequences without indels, from 85 clones with sizes ranging from a single sequence to 73 sequences. These sequences were taken from seven different samples from the dataset described above. The simulation ran on each clone 10 independent times, each time with different random variables as described above, in order to extend the dataset. These simulations yielded a total of 10,475 sequences, out of which 10,308 sequences had artifact deletions.

We then ran Ig-Indel-Identifier on each clone from the simulated dataset, using all possible combinations of the following program parameters: the minimum HPT length (with values ranging between 2 and 5) and the required number of sequences sharing the same indel for an indel to be considered a legitimate indel (with values ranging between 1 and 12). The former value range was based on our finding that there were no HPTs of length higher than five in our dataset.

For each parameter combination, we recorded how many artifact deletions were identified and calculated the percentage of accuracy (the number of identified deletions divided by the total number of artifact deletions and multiplied by 100).

As expected, the higher the value for HPT length, the lower the number of identified artifact deletions. This is reasonable since the probability of a HPT to be found in a sequence decreases with its length (HPTs of length two are much more frequent than HPTs of length five).

On the contrary, but also as expected, the higher the required number of sequences sharing the same deletion, the more artifact deletions were identified, designated as suspect and finally discarded. This is also reasonable, since the higher the required number of sequences sharing the same indel, the more stringent the requirements, and hence more indels (and sequences) do not fulfill these requirements.

Table 5 presents the average required number of sequences sharing the same deletion in each HPT length that was needed to identify 50% of the artifact deletions, or to identify all the artifact deletions—or as many as the program managed to identify. Due to the fact that Ig-Indel-Identifier was run individually on different samples (in our case we used seven different samples to collect the initial dataset of clones and sequences without indels from), and since each Ig-Indel-Identifier run included all combinations as explained above, the results in **Table 5** represent the average numbers out of 336 ($7 \times 4 \times 12$) runs.

When the minimal HPT length was 2, 50% of the artifact deletions were identified only when we required more than five sequences sharing the same deletion (on average) for a deletion to be considered legitimate. The maximal number of the artifact deletions identified out of the total deletions in the dataset was

Table 5 | The numbers of sequences sharing an indel within a clone that are required in order for it to be considered legitimate, that allow the indicated level of artifact indel identification.

HPT length	50% identification	Maximal identification
2	5	10
3	7	10
4	—	9*
5	—	7*

Numbers marked with a "" indicate that for the indicated HPT length, the program achieved less than 50% identification even with the indicated number of required sequences. Higher values gave the same results, so the minimal values of the required number of sequences in a clone that share the same indel were chosen.

obtained only when we required more than 10 sequences sharing the same deletion (on average). Similarly, when the HPT length was 3 (or 4), 50% of the artifact deletions were identified when we required more than 7 (or 9) sequences sharing the same deletion (on average). For HPT = 3, the maximal number of the artifact deletions identified out of the total deletions in the dataset was when we required more than 10 sequences sharing the same deletion (on average). With HPTs of length 4–5, no required number of sequences sharing the same indel could help identify more than 50% of the artifact indels. If the HPT length is set to 4 in Ig-Indel-Identifier, the program does not consider HPTs of length less than 4 and hence “misses” those indels, which brings the % identification down no matter how many sequences we require. When the HPT length was 5, the maximal number of the artifact deletions identified out of the total deletions in the dataset was when we required more than seven sequences sharing the same deletion (on average). Again, most of the deletions occurred near or inside HPTs of length less than 5, thus, no matter what the number of sequences sharing the same indel was, most of the indels were missed. Based on these results, our conclusion is that one should consider using either 2 or 3 as the values for HPT length in Ig-Indel-Identifier; otherwise the program would miss many artifact indels. Of course, it would be more efficient for each user to investigate their data for appearances of indels near or inside HPTs and lengths of the latter, in order to decide on the appropriate parameter values. We also recommend demanding as many sequences to share the same indel as possible (each user should optimize this number for their specific dataset).

In addition, we tested Ig-Indel-Identifier performance on 2355 Ig Sanger sequences from data published on B cells from autoimmune diseases (AI) (Zuckerman et al., 2010a,b) and lymphomas (Zuckerman et al., 2010c). The Sanger sequences barely contained

Table 6 | Numbers of Sanger and 454 sequences after Ig-Indel-Identifier cleaning.

Sequencing method	Sample (number of sequences in the sample)	Number of indels found in the sample	Number of sequences with artifact indels ^a	Number of uncertain indels ^a	Number of sequences w/o indels or with legitimate indels	Total number of point mutations in the sample	Number of point mutations in AID targeting motifs
Sanger	MG (33)	24	2	0	31	459	11
	MS (78)	4	0	0	78	2709	25
	Myositis-Bradshaw (33)	25	4	0	29	781	0
	RA-Gause (28)	7	1	2	25	356	6
	RA-Miura (123)	125	10	73	41	2495	29
	SS-Gellrich (70)	2	0	0	70	1269	13
	SS-Jacobi (190)	94	14	58	118	1226	22
	BL-Chapman (22)	0	0	0	22	465	1
	MZL-Zhu (72)	0	0	0	72	587	8
	DLBCL (708)	98	17	0	692	19,416	209
	FL (772)	87	31	0	741	19,249	226
	PCNSL (226)	4	2	0	224	7204	127
454	LN-1 (507)	2276	24	1	482	4631	2
	LN-2 (913)	8164	82	1	837	5682	0
	LN-3 (585)	5803	30	0	557	3019	0
	LN-4 (1137)	17,831	280	2	887	5744	0

^aSame as in **Table 4**.

MG, Myasthenia Gravis; MS, Multiple Sclerosis; RA, Rheumatoid Arthritis; SS, Sjögren's Syndrome; BL, Burkitt's Lymphoma; MZL, Marginal Zone Lymphoma; DLBCL, Diffuse Large B Cell Lymphoma; FL, Follicular Lymphoma; PCNSL, Primary Central Nervous System Lymphoma; LN, Lymph Node.

For the original studies from which the sequences were taken, see Zuckerman et al. (2010a,b,c).

indels (**Table 6**). However, when they did, most of the indels were uncertain, due to the small numbers of sequences sampled using the Sanger method. Only a small proportion of the sequences contained artifact indels. Therefore, the many indels observed in the 454 sequences and identified by Ig-Indel-Identifier are probably sequencing errors.

DISCUSSION

HTS is increasingly popular in various research fields such as immunology, cancer research, and evolutionary biology. The enormous amounts of data generated by HTS require the development of new and efficient data processing algorithms. There are already several tools for cleaning and analysis of HTS data, but no dedicated program for Ig genes has been made publicly available up to this work. In this paper, we present Ig-HTS-Cleaner, a program that successfully performs the pre-processing of Ig sequences derived from HTS, and Ig-Indel-Identifier, a program that precisely distinguishes between legitimate and artifact indels

which are typical of 454 HTS (and discards the latter), and also discards sequences containing point mutations with low quality score that appear only once in a clone. The two programs are independent of each other or any other tools, and are applicable to other sequences, in addition to Ig genes, and other sequencing platforms in addition to 454.

While the rules defined in Ig-Indel-Identifier do not guarantee that we identify all sequencing artifacts, it is known that HTS using the 454 platform mostly introduces indels near HPTs (Huse et al., 2007), while SHM of Ig genes mostly introduces point mutations rather than indels (Liu and Schatz, 2009; Steele, 2009). Moreover, most features of the SHM process are studied through analysis of point mutations (Zuckerman et al., 2010a,b,c). Thus, on one hand, the elimination of a large fraction of the artifact indels helps us retain the legitimate sequences that—having fewer indels—are easier to align and analyze further. On the other hand, the remaining artifact indels that may have not been eliminated do not affect the measurements of mutation characteristics.

The next step in analyzing Ig genes is to identify the GL V(D)J segments used in the unmutated (ancestor) sequences, and then assign the sequences into clonally-related groups. For gene segment identification, we use either SoDA (Volpe et al., 2006) or iHMMune-align (Gaëta et al., 2007). Our automated pipeline, which follows the above-described processes of cleaning the data, removing indels and identifying GL segments, contains our program “Ig_Clone_Finder[®],” which groups the sequences into clones based only on their V, D, and J segments. The weakness of this method is that two different rearrangements using the same V(D)J segments may be grouped together into the same clone; this necessitates manual checking of groups that clearly segregate into two or more different clones. A more sophisticated method, based on sequence clustering and the use of an empirical cut-off, was recently published by Chen et al. (2010); however, it has yet to be tested on large data sets.

Another weak point of HTS analysis is identifying chimeric (hybrid) sequences generated during PCR or HTS. It is essential to discard such sequences before performing repertoire and hypermutation analyses. Although there are a few existing tools that identify and discard chimeric sequences from HTS data (Huber et al., 2004), these programs are not suitable for use with Ig gene sequences. These programs depend on reference sequences that do not exist for Ig gene sequences, due to the complexity of Ig gene rearrangements and mutations, which make almost each sequence unique. There are several reasons why such a tool has not been fully developed yet for Ig genes, although HTS is already available and is extensively used. One major reason is the large homology between V segments. In order to find a chimeric sequence, one must recognize the two most probable

V segments [obtained by e.g., SoDA2 (Munshaw and Kepler, 2010) or iHMMune-Align (Gaëta et al., 2007)] that are likely to have been merged to create the suspected sequence. The problem is that GL genes from the same family can only diverge by up to 25% (by definition) and are usually much more similar, while mutated Ig genes can diverge by several tens of mutations (Cook and Tomlinson, 1995; Rajewsky, 1996). Thus, it is often impossible to identify whether the mismatches in the alignment are due to SHM of the suspected sequence, or due to SNPs that distinguish between the two almost identical V segments. Due to these reasons, we still search for chimeric sequences manually. However, discarding sequences that are too short or too long to be legitimate Ig gene sequences, as we do, probably gets rid of some obvious chimeras.

Technologies for HTS and the amounts of sequences generated using them continue to evolve. For this reason, we developed Ig-HTS-Cleaner and Ig-Indel-Identifier, two independent programs for cleaning high-throughput sequences. We hope these programs would be useful to the research community.

ACKNOWLEDGMENTS

The authors are indebted to Dr. Deborah Dunn-Walters for critical reading of the manuscript. This work was supported in parts by an Israel Science Foundation [grant number 270/09, to Ramit Mehr and Iris Barshack]; and a Human Frontiers Science Program Research Grant [to Ramit Mehr]. The work was part of Miri Michaeli’s studies toward the MSc degree in Bar-Ilan University, and she was supported by a Combined Technologies Scholarship from the Israeli Council for Higher Education.

REFERENCES

- Ademokun, A., Wu, Y.-C., and Dunn-Walters, D. K. (2010). The ageing B cell population: composition and function. *Biogerontology* 11, 125–137.
- Albers, C. A., Lunter, G., Macarthur, D. G., McVean, G., Ouwehand, W. H., and Durbin, R. (2011). Dindel: accurate indel calls from short-read data. *Genome Res.* 21, 961–973.
- Aronesty, E. (2011). ea-utils: Command-line Tools for Processing Biological Sequencing Data. Available online at: <http://code.google.com/p/ea-utils>
- Bansal, V., and Libiger, O. (2011). A probabilistic method for the detection and genotyping of small indels from population-scale sequence data. *Bioinformatics* 27, 2047–2053.
- Barak, M., Zuckerman, N. S., Edelman, H., Unger, R., and Mehr, R. (2008). IgTree: creating Immunoglobulin variable region gene lineage trees. *J. Immunol. Methods* 338, 67–74.
- Blanca, J. M., Pascual, L., Ziarsolo, P., Nuez, F., and Cañizares, J. (2011). ngs _ backbone: a pipeline for read cleaning, mapping and SNP calling using next generation sequence. *BMC Genomics* 12:285. doi: 10.1186/1471-2164-12-285
- Bolger, A., and Giorgi, F. Trimmomatic: A Flexible Read Trimming Tool for Illumina NGS Data. Available online at: <http://www.usadellab.org/cms/index.php?page=trimmomatic>
- Boyd, S. D., Gaëta, B. A., Jackson, K. J. L., Fire, A. Z., Marshall, E. L., Merker, J. D., et al. (2010). Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* 184, 6986–6992.
- Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., et al. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. *Sci. Transl. Med.* 1, 12ra23.
- Buffalo, V. Scythe. Available online at: <https://github.com/vsbuffalo/scythe>
- Campbell, P. J., Pleasance, E. D., Stephens, P. J., Dicks, E., Rance, R., Goodhead, I., et al. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 105, 13081–13086.
- Chen, Z., Collins, A. M., Wang, Y., and Gaëta, B. A. (2010). Clustering-based identification of clonally-related immunoglobulin gene sequence sets. *Immunome Res.* 6(Suppl. 1), S4.
- Cook, G. P., and Tomlinson, I. M. (1995). The human immunoglobulin VH repertoire. *Immunol. Today* 16, 237–242.
- Dalca, A. V., Rumble, S. M., Levy, S., and Bradno, M. (2010). VARiD: a variation detection framework for color-space and letter-space platforms. *Bioinformatics* 26, i343–i349.
- Dunn-Walters, D. K., and Ademokun, A. (2010). B cell repertoire and ageing. *Curr. Opin. Immunol.* 22, 514–520.
- Falgueras, J., Lara, A. J., Fernández-Pozo, N., Cantón, F. R., Pérez-Trabado, G., and Claros, M. G. (2010). SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics* 11:38. doi: 10.1186/1471-2105-11-38
- Gaëta, B. A., Malming, H. R., Jackson, K. J. L., Bain, M. E., Wilson, P., and Collins, A. M. (2007). iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* 23, 1580–1587.
- Galan, M., Guivier, E., Caraux, G., Charbonnel, N., and Cosson, J.-F. (2010). A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics* 11:296. doi: 10.1186/1471-2164-11-296
- Gibson, K. L., Wu, Y.-C., Barnett, Y., Duggan, O., Vaughan, R., Kondeatis, E., et al. (2009). B-cell diversity decreases in old age and is correlated with poor health status. *Aging cell* 8, 18–25.
- Gordon, A. FASTX-Toolkit. Available online at: <http://hannonlab.cshl.edu/fastxtoolkit/>
- Huber, T., Faulkner, G., and Hugenholtz, P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence

- alignments. *Bioinformatics* 20, 2317–2319.
- Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., and Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8, R143–R151.
- John, J. S. *SeqPrep*. Available online at: <https://github.com/jstjohn/SeqPrep>
- Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., et al. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25, 2283–2285.
- Kong, Y. (2011). Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* 98, 152–153.
- Krueger, F. *Trim Galore!* Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/trimgalore/>
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948.
- Lassmann, T., Hayashizaki, Y., and Daub, C. O. (2009). TagDust—a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* 25, 2839–2840.
- Lindgreen, S. (2012). *AdapterRemoval: Easy Cleaning of Next Generation Sequencing Reads*. Available online at: <http://code.google.com/p/adapterremoval/>
- Liu, M., and Schatz, D. G. (2009). Balancing AID and DNA repair during somatic hypermutation. *Trends Immunol.* 30, 173–181.
- MacCarthy, T., Kalis, S. L., Roa, S., Pham, P., Goodman, M. F., Scharff, M. D., et al. (2009). V-region mutation *in vitro*, *in vivo*, and *in silico* reveal the importance of the enzymatic properties of AID and the sequence environment. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8629–8634.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12.
- Morgan, M., Anders, S., Lawrence, M., Aboyou, P., Pagès, H., and Gentleman, R. (2009). ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* 25, 2607–2608.
- Munshaw, S., and Kepler, T. B. (2010). SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements. *Bioinformatics* 26, 867–872.
- Pandey, R. V., Nolte, V., and Schlötterer, C. (2010). CANGS: a user-friendly utility for processing and analyzing 454 GS-FLX data in biodiversity studies. *BMC Res. Notes* 3:3. doi: 10.1186/1756-0500-3-3
- Rajewsky, K. (1996). Clonal selection and learning in the antibody system. *Nature* 381, 751–758.
- Scheid, J. F., Mouquet, H., Feldhahn, N., Seaman, M. S., Velinzon, K., Pietzsch, J., et al. (2009). Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature* 458, 636–640.
- Schmieder, R., Lim, Y. W., Rohwer, F., and Edwards, R. (2010). TagCleaner: identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* 11:341. doi: 10.1186/1471-2105-11-341
- Steele, E. J. (2009). Mechanism of somatic hypermutation: critical analysis of strand biased mutation signatures at A:T and G:C base pairs. *Mol. Immunol.* 46, 305–320.
- Tabibian-Keissar, H., Zuckerman, N. S., Barak, M., Dunn-Walters, D. K., Steiman-Shimony, A., Chowers, Y., et al. (2008). B-cell clonal diversification and gut-lymph node trafficking in ulcerative colitis revealed using lineage tree analysis. *Eur. J. Immunol.* 38, 2600–2609.
- Unknown. *FAR – The Flexible Adapter Remover*. Available online at: <http://sourceforge.net/apps/mediawiki/theflexibleadap/>
- Volpe, J. M., Cowell, L. G., and Kepler, T. B. (2006). SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics* 22, 438–444.
- Wu, Y.-C., Kipling, D., Leong, H. S., Martin, V., Ademokun, A., and Dunn-Walters, D. K. (2010). High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* 116, 1070–1078.
- Zuckerman, N. S., Hazanov, H., Barak, M., Edelman, H., Hess, S., Shcolnik, H., et al. (2010a). Somatic hypermutation and antigen-driven selection of B cells are altered in autoimmune diseases. *J. Autoimmun.* 35, 325–335.
- Zuckerman, N. S., Howard, W. A., Bismuth, J., Gibson, K. L., Edelman, H., Berrih-Aknin, S., et al. (2010b). Ectopic GC in the thymus of myasthenia gravis patients show characteristics of normal GC. *Eur. J. Immunol.* 40, 1150–1161.
- Zuckerman, N. S., McCann, K. J., Ottensmeier, C. H., Barak, M., Shahaf, G., Edelman, H., et al. (2010c). Immunoglobulin gene diversification and selection in follicular lymphoma, diffuse large B cell lymphoma and primary central nervous system lymphoma revealed by lineage tree and mutation analyses. *Int. Immunol.* 22, 875–887.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 July 2012; **paper pending published:** 05 September 2012; **accepted:** 30 November 2012; **published online:** 28 December 2012.

Citation: Michaeli M, Noga H, Tabibian-Keissar H, Barshack I and Mehr R (2012) Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing. *Front. Immunol.* 3:386. doi: 10.3389/fimmu.2012.00386

This article was submitted to Frontiers in B Cell Biology, a specialty of Frontiers in Immunology.

Copyright © 2012 Michaeli, Noga, Tabibian-Keissar, Barshack and Mehr. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



Immunoglobulin Analysis Tool: a novel tool for the analysis of human and mouse heavy and light chain transcripts

Tobias Rogosch¹, Sebastian Kerzel¹, Kam Hon Hoi², Zhixin Zhang³, Rolf F. Maier¹, Gregory C. Ippolito⁴ and Michael Zemlin^{1*}

¹ Department of Pediatrics, Philipps-University Marburg, Marburg, Germany

² Department of Biomedical Engineering, University of Texas at Austin, Austin, TX, USA

³ Department of Pathology and Microbiology, Eppley Institute for Research in Cancer, University of Nebraska Medical Center, Omaha, NE, USA

⁴ Section of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, TX, USA

Edited by:

Harry W. Schroeder, University of Alabama at Birmingham, USA

Reviewed by:

John D. Colgan, University of Iowa, USA

Deborah K. Dunn-Walters, King's College London School of Medicine, UK

***Correspondence:**

Michael Zemlin, University Children's Hospital, Baldingerstrasse, D-35033 Marburg, Germany.

e-mail: zemlin@med.uni-marburg.de

Sequence analysis of immunoglobulin (Ig) heavy and light chain transcripts can refine categorization of B cell subpopulations and can shed light on the selective forces that act during immune responses or immune dysregulation, such as autoimmunity, allergy, and B cell malignancy. High-throughput sequencing yields Ig transcript collections of unprecedented size. The authoritative web-based IMGT/HighV-QUEST program is capable of analyzing large collections of transcripts and provides annotated output files to describe many key properties of Ig transcripts. However, additional processing of these flat files is required to create figures, or to facilitate analysis of additional features and comparisons between sequence sets. We present an easy-to-use Microsoft® Excel® based software, named Immunoglobulin Analysis Tool (IgAT), for the summary, interrogation, and further processing of IMGT/HighV-QUEST output files. IgAT generates descriptive statistics and high-quality figures for collections of murine or human Ig heavy or light chain transcripts ranging from 1 to 150,000 sequences. In addition to traditionally studied properties of Ig transcripts – such as the usage of germline gene segments, or the length and composition of the CDR-3 region – IgAT also uses published algorithms to calculate the probability of antigen selection based on somatic mutational patterns, the average hydrophobicity of the antigen-binding sites, and predictable structural properties of the CDR-H3 loop according to Shirai's *H3-rules*. These refined analyses provide in-depth information about the selective forces acting upon Ig repertoires and allow the statistical and graphical comparison of two or more sequence sets. IgAT is easy to use on any computer running Excel® 2003 or higher. Thus, IgAT is a useful tool to gain insights into the selective forces and functional properties of small to extremely large collections of Ig transcripts, thereby assisting a researcher to mine a data set to its fullest.

Keywords: immunoglobulin heavy chain gene, immunoglobulin light chain gene, rearrangement, somatic mutation, sequence analysis software, antibody repertoire, high-throughput analysis, deep sequencing

INTRODUCTION

The fate of a B cell largely depends on the B cell receptor, or immunoglobulin (Ig), which it expresses on its surface (Rajewsky, 1996; Kurosaki et al., 2010). Thus, the analysis of Ig gene transcripts can give important insights into the selective forces that act upon B cells during cellular maturation or during physiological or pathological immune reactions (Schroeder and Cavacini, 2010). For example, repertoire studies of Ig transcripts have revealed that the length and composition of the Ig heavy chain third complementarity determining region (CDR-H3) is strictly regulated during ontogeny, and somatic mutations are rare during the perinatal period even in secondary antibody repertoires (Schroeder et al., 1987, 2001; Cuisinier et al., 1993; Brezinschek et al., 1997; Zemlin et al., 2001, 2007; Kolar et al., 2004; Souto-Carneiro et al., 2005; Schelonka et al., 2007; Richl et al., 2008; Prabakaran et al., 2012). It has also been shown that the composition of the antigen-binding site plays a key role during B cell maturation and during

the recruitment into various B cell subsets (Schelonka et al., 2007; Arnaout et al., 2011) and during protective immune responses (Rajewsky, 1996; Frolich et al., 2010). Moreover, studies of Ig repertoires can give valuable insights into the immune dysregulation that underlies the development of autoimmunity (Dorner and Lipsky, 2005; Vrolix et al., 2010; Zuckerman et al., 2010; Kalininina et al., 2011) and allergies (Snow et al., 1998; Takhar et al., 2007; Kerzel et al., 2010).

The antigen-binding site of the antibody is endowed with an almost unlimited theoretical diversity due to the imprecise junction of Variable, Joining, and (in the case of the Ig heavy chain) Diversity gene segments (Tonegawa, 1983). The random exonucleolytic truncation of the rearranged gene segments and the insertion of non-encoded N-nucleotides and P-nucleotides, the shuffling of light and heavy chains, and the insertion of somatic mutations during the germinal center reaction further expands the potential diversity exponentially. Theoretically, these mechanisms

Table 1 | Estimate of the maximum size of sequence collections that can be processed.

Excel version	Operation system	Max. memory	No. of sequences
Excel 2003	Windows XP	1 Gigabyte	~40,000
	Windows 7 (32/64-bit)		
Excel 2007	Windows XP	2 Gigabyte	~60,000
Excel 2010 (32-bit)	Windows 7 (32/64-bit)		
Excel 2010 (64-bit)	Windows 7 (64-bit)	8 Terabyte	150,000 (max. no. of IMGT/HighV-QUEST)

The restrictions are caused by limited addressable memory by Excel. Excel versions prior 2007 can not address more than 1 GB of memory. 32-bit versions of Excel 2007/2010 can use 2 GB of memory, while the 64-bit versions are virtually unrestricted.

allow the production of more than 10^{15} different antigen-binding sites (Schroeder and Cavacini, 2010). Although seemingly limitless in theoretical potential, the human antibody response probably does not exploit more than 1% of its potential diversity (Boyd et al., 2009; Glanville et al., 2009; Arnaout et al., 2011). Thus, it seems unlikely that the expressed antibody repertoire would represent merely a random selection of the theoretical diversity.

In order to discover potential biases within repertoires that may have been coined by selective forces, it is desirable to study large numbers of Ig gene transcripts. With the advent of next generation sequencing (NGS) technologies, such as Roche 454 pyrosequencing, the direct large-scale sampling of sequence collections of 10^4 , 10^5 , and even greater numbers, is now obtainable within the span of a few days (Boyd et al., 2009; Reddy et al., 2010; Wu et al., 2010; Zuckerman et al., 2010; Jiang et al., 2011; Ippolito et al., 2012). Previously published semi-automated instruments cannot be used for such large collections or state-of-the-art characterizations due to significant quantitative and qualitative advances in Ig gene analysis (Shannon, 1997; Johnson and Wu, 2000; Zemlin et al., 2003). Thus, novel analysis tools are required which can handle extremely large sequence batches.

The online repository “International ImMunoGeneTics Information System®” (IMGT®¹, founder and director: Marie-Paule Lefranc, Montpellier, France (Brochet et al., 2008; Lefranc et al., 2009) offers IMGT/HighV-QUEST, a free online tool to assign Variable, Diversity, and Joining gene segments to each individual full-length Ig transcript in batches up to 150,000 sequences. In addition, IMGT/HighV-QUEST provides numerous descriptors for each individual sequence, such as assignment of N- and P-nucleotides, amino acid translation, position of somatic mutations, isoelectric point, and many others (Giudicelli et al., 2011; Alamyar et al., 2012). The output files of these analyses contain descriptions of each individual sequence and can be downloaded as text files in comma separated values (CSV) format for documentation and further analysis.

Our aim was to create an easy-to-use software tool for the generation of informative statistics and publication-ready figures derived from the HighV-QUEST text-only output files. Moreover, we sought to include new and important analyses of higher order antibody features. For instance, although Shirai's *H3-rules* have been formulated for the sequence-based prediction of CDR-H3 structural properties (Shirai et al., 1999), and whereas complex algorithms have been published to determine the probability by which a somatic mutation profile might arise non-randomly from antigen-driven selection (Chang and Casali, 1994; Losso et al., 2000), there are at present no software tools available to the research community for high-throughput application of these rules and algorithms.

Here we present Immunoglobulin Analysis Tool (IgAT), a novel and user-friendly software tool for the extensive analysis and graphical presentation of very large collections of Ig transcripts which have been pre-analyzed by IMGT/HighV-QUEST. IgAT additionally calculates the probability of antigen-driven selection within Ig repertoires and predicts structural properties of the antigen-binding site. IgAT can be used to analyze up to 150,000 human or murine heavy or light chain transcripts in a single run of the application and automatically generates 25 Microsoft® PowerPoint® graphics files illustrating key characteristics of the Ig repertoire, such as VDJ gene utilization, amino acid use, CDR-H3 junctional diversity, and average hydrophobicity, as well as the quantitation of somatic mutation among Ig heavy chain transcripts, to name but a few. IgAT is available free of charge.

When applied to two or more sequence collections (e.g., samples from multiple individuals, different cell subsets, or identical cell subsets but under differing immunological conditions), IgAT readily yields the necessary data to allow statistical and graphical comparisons between various repertoires.

METHODS

IgAT is a Microsoft® Excel® workbook containing the analysis functions as Visual Basic® for Applications (VBA) code. Each sheet is described in the results section. The workbook was created in Excel 2010 on Microsoft Windows® XP but should be compatible with Excel versions down to Excel 2003 with some limitations (Table 1). IgAT is not compatible with Excel for Mac®. The file can be found at: www.uni-marburg.de/neonat/igat

RESULTS

In the following, we present the features offered by IgAT, using exemplarily a previously published collection of 78,569 murine Ig heavy chain sequences that contained 18,403 functional sequences (Reddy et al., 2010). These sequences were obtained from CD138⁺ plasma-cell-enriched bone marrow mRNA of two BALB/c mice immunized with human complement serine protease (C1S; NCBI Entrez Gene ID: 716).

Begin with a text file of FASTA-formatted Ig DNA sequences as can be obtained from a Roche 454 experimental run or other techniques. When submitting the sequence batch to IMGT/HighV-QUEST, under the advanced parameters setting, “Nb of accepted D-GENE in JUNCTION” must be set to the default (1) as IgAT will only process IMGT output files that assign a maximum of one

¹<http://www.imgt.org>

single D-gene to each V-D_H-J junction. IMGT individual result files are not necessary for the analysis with IgAT.

INPUT

As input, IgAT takes the 11 CSV text output files standardly generated by IMGT/HighV-QUEST derived from its analysis of raw 454 sequence data uploaded by the researcher. IgAT imports the folder containing the IMGT/HighV-QUEST CSV text output files through the cell “C6” of the “input” worksheet. (Alternatively, the IgAT program may be copied and pasted into the folder, which already contains the IMGT files.) Optionally, sequences marked as “unproductive” by IMGT/HighV-QUEST can be deleted. Deleting unproductive sequences will improve performance but might discard functional transcripts as Roche 454 sequencing is prone to homopolymer errors due to technical reasons.

The species (human or mouse), the Ig chain (heavy, lambda, or kappa), the minimum number of non-mutated nucleotides that are required to identify a diversity (D) gene, and the option to calculate the Taq-error must be chosen before starting the analysis. The Ig isotype is needed to calculate the Taq-error (**Figure 1**).

To start the analysis simply press the button “analyze data.” If “convert formulas to text” is checked, most formulas will be replaced by their values, resulting in reduced file size and recalculation time. In this case, however, additional changes will not have any effect on the analysis output. Once the sequence analysis is complete, the graphs can be exported as Microsoft PowerPoint® files (.ppt) by pressing “save graphs as ppt.”

The workbook was created in Excel 2010 and tested in Excel 2003 and 2010. To determine if your Microsoft Office® software meets this requirement, press “check office version.” It might be compatible with other versions (not tested).

SUMMARY

The number of total, non-functional, functional, and unique sequences, as well as the number of clonotypes is listed in the “summary” worksheet (**Figure 2**). Deep sequencing technologies usually yield a significant proportion of incomplete or otherwise defective sequences. IgAT counts the sequences which were labeled “unproductive,” “no result,” or “unknown” by IMGT/HighV-QUEST.

Sequences are considered clonally related if they (i) use the same V and J genes, (ii) have an identical CDR-3 length, and (iii) a highly homologous CDR-3 region. The default definition of “highly homologous CDR-3 region” is $\leq 10\%$ difference in nucleotide sequence. IgAT gives the user the flexibility to choose another percentage difference in nucleotide sequence, or a total number of nucleotide matches, or a percentage or total number difference in amino acid sequence when defining clonotypic parameters.

DATA

The “Data” worksheet contains the imported data of the IMGT/HighV-QUEST output files. IgAT uses the taxonomy and numbering of the IMGT repository (Lefranc et al., 2009).

IgAT
(Immunoglobulin Analysis Tool)

uses output file of IMGT®/HighV-QUEST version: 1.1.1
(Alamya, E. et al. (2012) *Immune Research* 8:12)

description <- optional

folder of IMGT files <- if empty, the folder containing this spreadsheet is selected

species <- select species (human or mouse)

chain <- select chain (heavy, lambda or kappa)

min. number of non mutated nt in D <- the no. of non mutated nt that are required to identify a diversity (D) gene (not functional, if a light chain is selected) (default: 6)

calculate Taq-error <- not recommended, if sequences do not contain constant region

Microsoft Excel™ 2003 or above is needed to do the analysis.
Microsoft Powerpoint™ 2003 or above is needed to save the graphs as ppt.

Press "Check Office Version" to check if your software meets the requirements.

Press "analyse data" to start analysis.

convert formulas to text

v1.00
by Tobias Rogosch & Michael Zemlin

INDEX

- [summary](#)
- [IMGT data](#)
- [sequences](#)
- [VDJ usage](#)
- [CDR3 length & composition](#)
- [somatic mutations & antigen selection](#)
- [AA usage \(CDR3\)](#)
- [AA frequency \(CDR3\)](#)
- [Kyte-Doolittle hydrophobicity \(CDR3\)](#)
- [reading frame usage \(D-gene\)](#)
- [prediction of CDR3 structure \(Shirai\)](#)
- [Taq-error](#)

FIGURE 1 | Screenshot of the “input” worksheet.

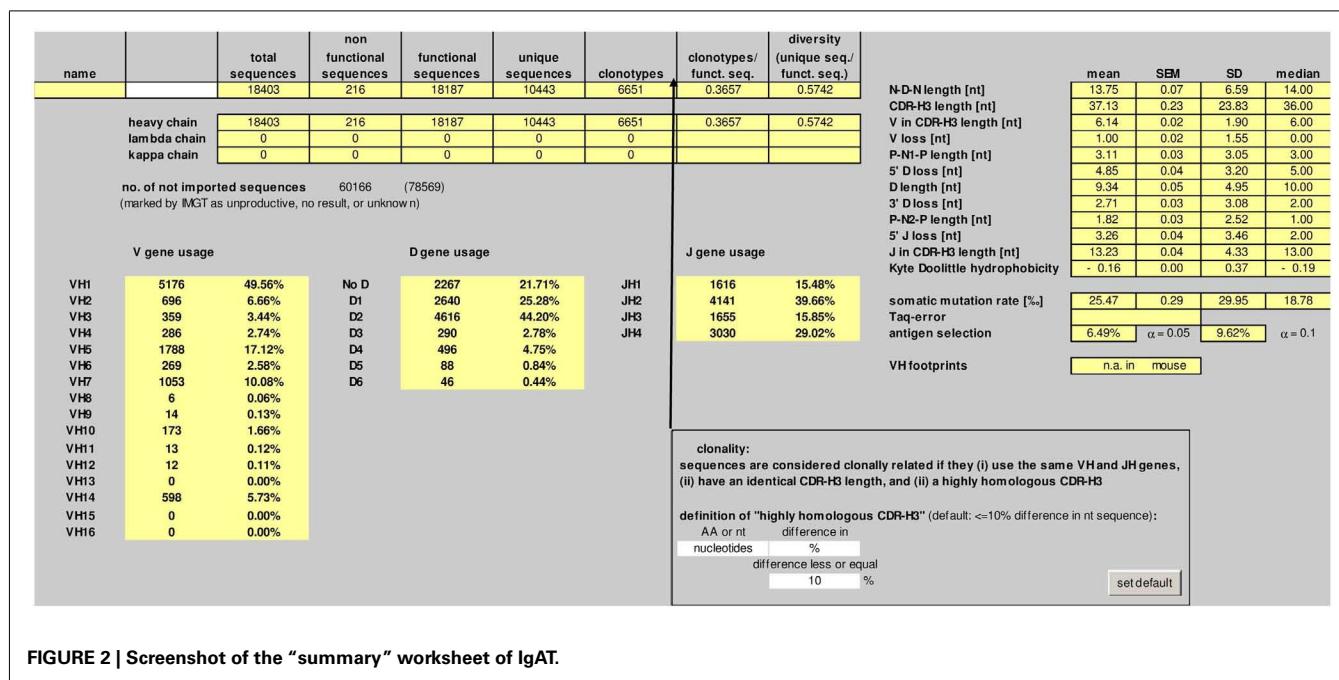


FIGURE 2 | Screenshot of the “summary” worksheet of IgAT.

SEQUENCE

In this worksheet, each nucleotide sequence occurs in an individual row and is split into framework regions (FR) 1–4 and complementarity determining regions 1–3. The sequences are ordered by functionality, which is defined by the existence of an open reading frame throughout the sequence, and by V gene segment utilization. Furthermore, the “Sequence” worksheet provides the length and amino acid translation for CDR-3, number of clonotypes, and identifies sequences with potential “V_H-replacement footprints” (only human sequences) that can originate from V_H replacement during receptor editing according to Zhang et al. (2003). In addition, sequences can be tagged with the sample ID. Based on sample IDs, the analysis can be confined to one or several samples or the transcripts can be divided into two groups for comparison.

VDJ

The “VDJ” worksheet contains absolute numbers, percentages, and graphs of the V-, D_H-, and J-gene families and individual genes in the order of their localization in the germline (Figure 3).

CDR-3_LENGTH

The “CDR-3_length” worksheet displays the nucleotide length distribution of CDR-3, N1-, and N2-nucleotides within the analyzed sequence collection (Figure 4). In addition, the average lengths of the components of CDR-3, namely V length, P-nucleotides 3' of V, N1-nucleotides, P-nucleotides 5' of D, D length, P-nucleotides 3' of D, N2-nucleotides, P-nucleotides 3' of J, and J length are displayed in a deconstruction graph. A separate graph displays the deconstruction of those sequences without an identifiable D-gene. As a default for the IgH chain, CDR-H3 is defined as amino acids 105–117, according to the IMGT unique numbering system. The descriptive statistics given in the “CDR_length” worksheet can be used for comparative statistics with other sequence collections.

SOMAT MUT

This worksheet displays the somatic mutation rate of each transcript (mutations per 1,000 nt), as well as the average mutational frequency (Figure 5A). In addition, the probability of antigen selection is analyzed by assessing the distribution of replacement and silent mutations between FRs and CDRs (only available for heavy chains). Using the method of Lossos et al. (2000), we determined the replacement frequency and the relative length of FR and CDR of each germline V_H gene. The average probability that a random mutation would allocate in CDR was calculated to be 0.23 ± 0.012, and the sequence-inherent probability that a mutation in the CDR would be a replacement mutation was estimated to be 0.79 ± 0.01. Therefore, the chance for a random mutation to introduce a replacement mutation into the CDR was 0.18. The binomial distribution method of Chang and Casali (1994) was used to calculate the 90 and 95% confidence limits for the ratio of replacement mutations in the CDR (R_{CDR}) to the number of total mutations in the V region (M_V) as described by Dahlke et al. (2006). These confidence intervals are shown as dark (90%) and light gray (95%) shaded area in Figure 5B. A data point falling outside these confidence limits represents a sequence that has a high proportion of replacement mutations in the CDR. Therefore, an allocation above the upper or below the lower confidence limit is considered indicative of Ag-driven selection. It should be mentioned that refined methods for calculation of antigen selection have been published and are available to the public (Hershberg et al., 2008; Uduman et al., 2011). However, at the present IgAT is not suitable to include this type of analyses, because sequence alignments in large sequence collections would require a different software environment.

AA

This worksheet shows the amino acid distribution and frequency of the CDR-3 loop for sequences with the same CDR-3 length as

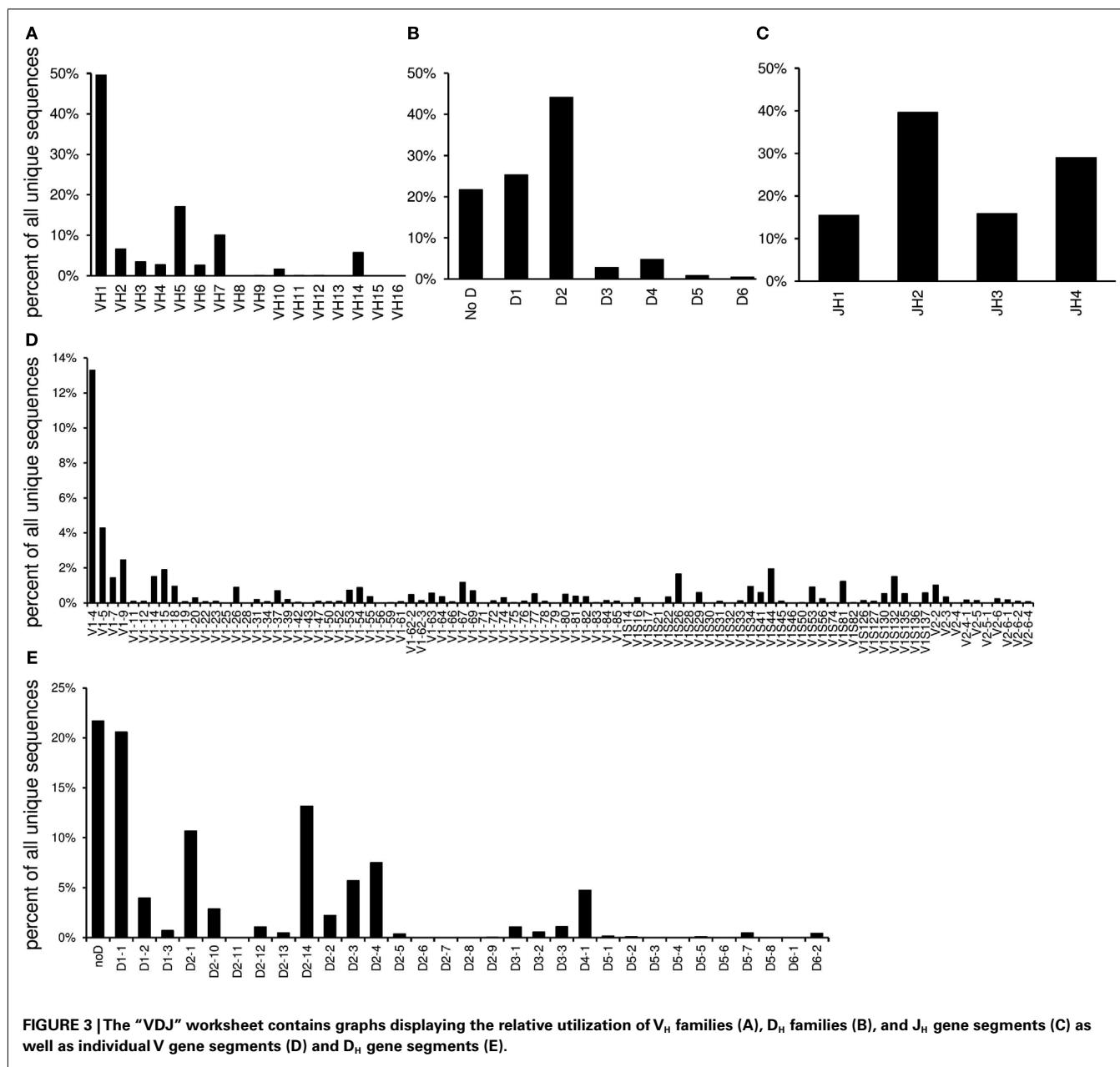


FIGURE 3 |The “VDJ” worksheet contains graphs displaying the relative utilization of V_H families (A), D_H families (B), and J_H gene segments (C) as well as individual V gene segments (D) and D_H gene segments (E).

entered in cell “G3” and different resulting amino acid variability plot (Shannon entropy, a measure of amino acid variability at a given position of aligned protein sequences, and Kabat–Wu plot, the number of different amino acids observed at a position divided by the frequency of the most common amino acid; Shannon, 1997; Johnson and Wu, 2000; Zemlin et al., 2003; Figure 6).

AA FREQUENCY

This diagram shows the amino acid frequencies of the CDR-3 loop for all sequences (Figure 7). The frequency is given as percent of all amino acids encoded by CDR-3 from all unique sequences studied. As a default for the IgH chain, the CDR-H3 loop is defined as the amino acids 107–114, according to the IMGT unique numbering

system, but the definition of the loop can be modified by the user by entering the limits into the worksheet “AA,” cells N5 and N6.

KYTE–DOOLITTLE

The normalized Kyte–Doolittle scale assigns one value to each amino acid. Negative numbers represent polar/hydrophilic amino acids and positive values represent hydrophobic amino acids (Kyte and Doolittle, 1982; Eisenberg, 1984). Figure 8 displays the distribution of average CDR-3 hydrophobicities according to the normalized Kyte–Doolittle scale.

IGHD

This worksheet displays the D_H gene reading frame usage (Figure 9). For each D_H segment there is one reading frame

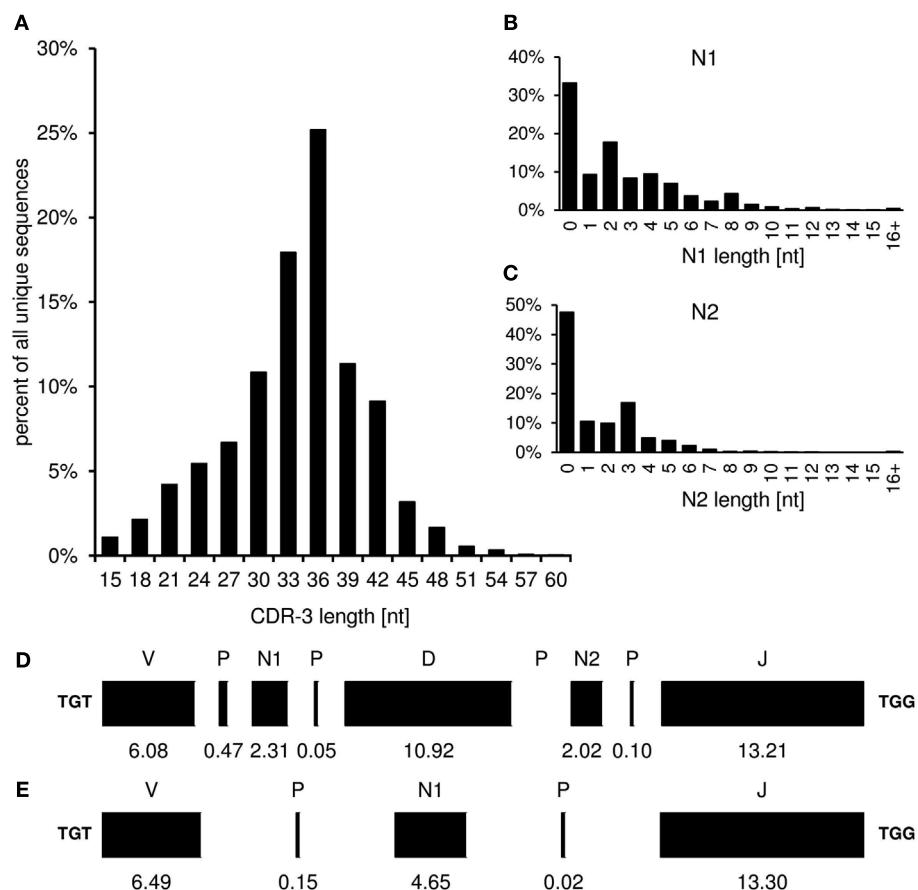


FIGURE 4 |The graphs in the “CDR-3_length” (positions 105–117) worksheet display the length distribution of CDR-H3 (A), N1 (B), N2 (C), and deconstruction graphs for CDR-H3 with (D) or without (E) identifiable D_H gene segment. Lengths are given in nucleotides.

encoding predominantly hydrophilic residues (especially tyrosine and serine; RF1), followed by a hydrophobic reading frame (RF2), and lastly a third reading frame that often encodes a stop codon (RF3). Thus, the third reading frame can be used only if either somatic mutations or else nucleotide losses during VDJ recombination delete the germline stop codon.

SHIRAI

In this worksheet the predicted structural features of the CDR-H3 are displayed (Figure 10). The “H3-rules” by Shirai (Shirai et al., 1999; Kuroda et al., 2008) are used to predict the structure of the CDR-H3 loop and base classified upon amino acid sequence, localization, and characteristics like hydrophobicity and size of the amino acid side chain. The structure of the base can be either extended, kinked, or extra kinked. In case of the latter two, the H3-rules may predict whether an intact hydrogen bond ladder or a deformed hairpin is formed within the loop.

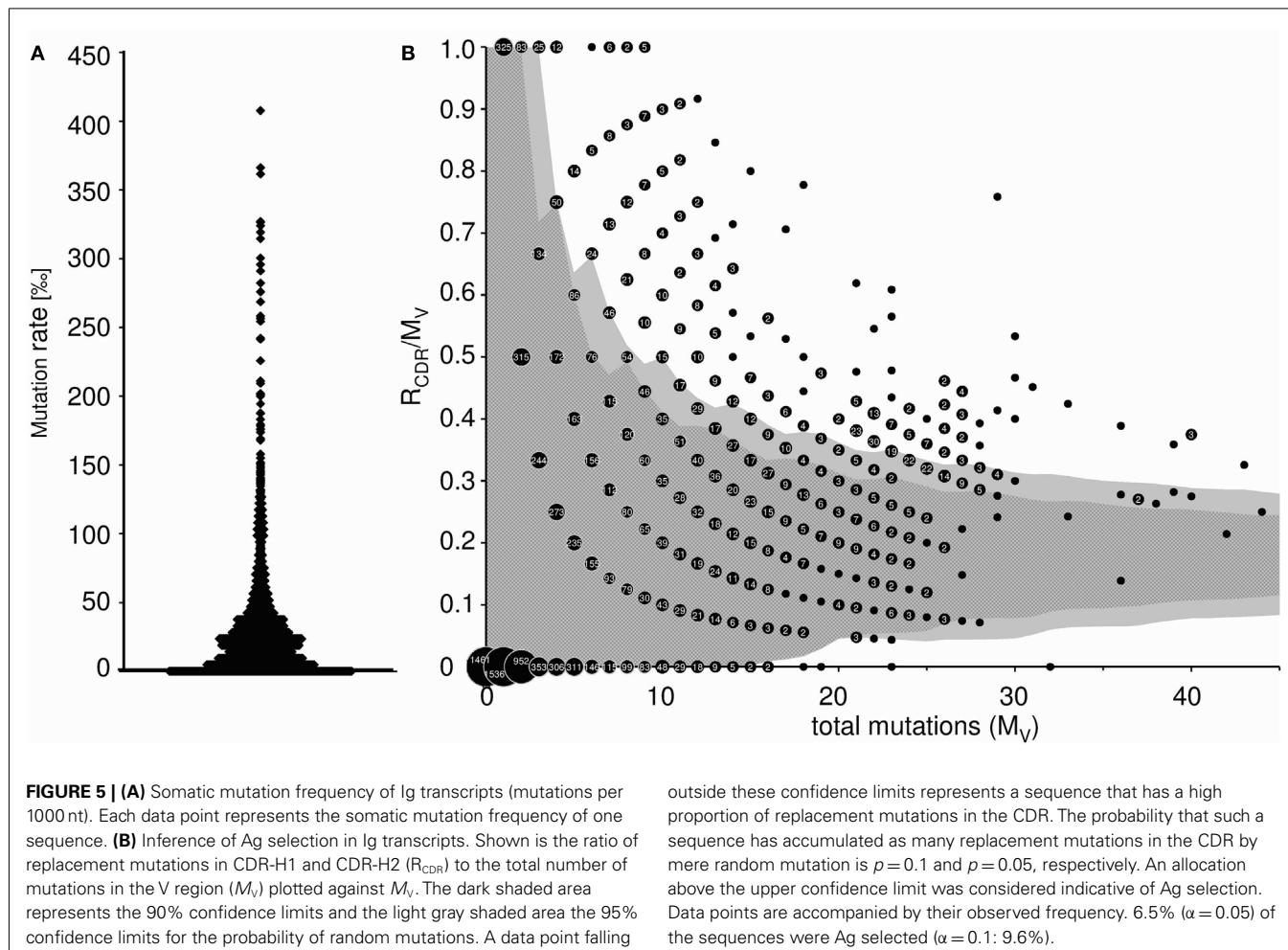
TAQ-ERROR

This worksheet calculates the Taq-error rate. To exclude a relevant biasing of the somatic mutation frequency by Taq polymerase errors, IgAT calculates the Taq-error rate within the stretches of the Ig constant region when it is included in the PCR amplicates.

DISCUSSION

Since the discovery of the Ig genes, as well as the fundamental mechanisms describing their combinatorial somatic rearrangement, numerous studies have been published with the goal of understanding the selective forces which might govern B cell and T cell development and the diversification of their lymphocyte receptor repertoires. Whereas B and T cells share a common mode of initial diversification (VDJ recombination), it is only B cells which include additional postrecombination diversification mechanisms such as V_H replacement and somatic hypermutation. Furthermore, whereas the selective forces shaping the receptor repertoire of developing T cells have been well established (Morris and Allen, 2012), the same cannot be said for the antibody receptor repertoire of B cells. For instance, mechanisms of positive selection are not clearly defined for B cell antibody repertoires; however, on the contrary, there are clear examples of negative selective mechanisms (deletion, anergy, and follicular exclusion) as well as additional mechanisms (average amino acid hydrophobicity of CDR-H3, preferential V gene utilization, V_H gene replacement) which act to constrain the diversity of the antibody repertoire.

Early pioneering efforts involved laborious cloning and classic Sanger DNA/cDNA sequencing which yielded sequence collections of modest size on the order of tens to a few hundreds.



Novel antibody repertoire studies employ high-throughput deep sequencing technologies which can yield collections of unprecedented sizes on the order of thousands to millions of raw sequence reads (reviewed in Benichou et al., 2011). To facilitate such studies, the web-based IMGT/HighV-QUEST™ program is capable of analyzing large collections of transcripts (up to 150,000 per analysis) by comparison with the known V, D_H, and J germline gene segments. Here we present IgAT, a novel easy-to-use Microsoft Excel based Visual Basic code for the summary, interrogation, and further processing of IMGT/HighV-QUEST output files. IgAT presents the data as organized spreadsheets, yields ready-to-publish statistics and figures, and allows the standardized comparison of multiple sequence batches.

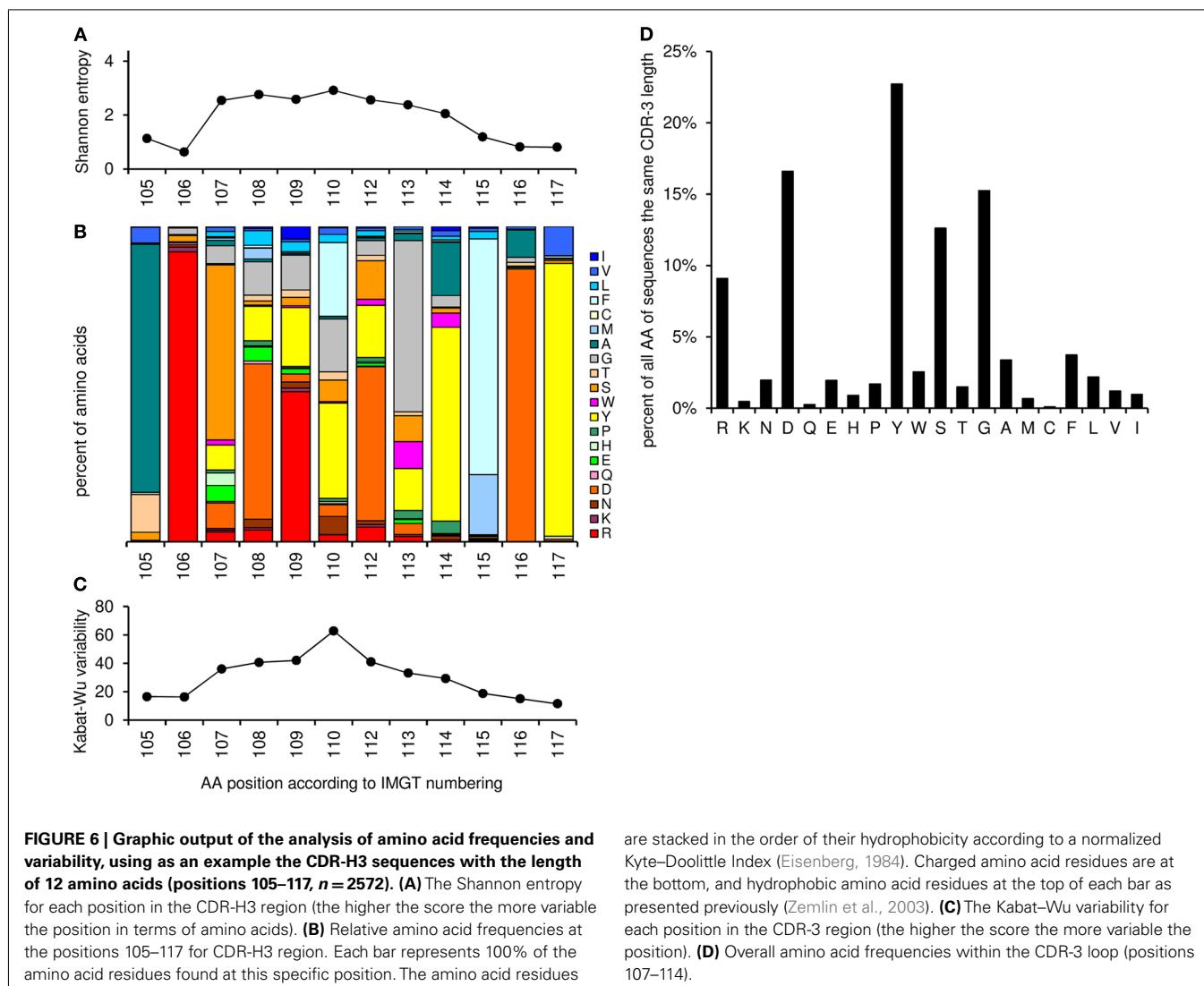
Conventional and Roche 454 deep sequencing of Ig heavy chain transcripts has been used to better understand the maturation of B cells, their selection into various maturational subsets (Wu et al., 2010), to determine the degree to which the repertoire might be genetically predetermined (Glanville et al., 2009; Ippolito et al., 2012), to characterize protective antibody responses (e.g., tetanus-toxoid neutralizing antibodies (Frolich et al., 2010) or HIV neutralizing antibodies (Wu et al., 2011), autoimmunity (Dorner and Lipsky, 2005; Vrolix et al., 2010; Zuckerman et al., 2010; Kalinina et al., 2011), allergies (Kerzel et al., 2010), and

especially a push to monitor minimal residual disease in B cell neoplasias (Boyd et al., 2009; Logan et al., 2011). In such studies, IgAT could help indicate to what extent the repertoire has been influenced by antigen-driven selection. The detailed analyses provided by IgAT can be used to speculate about the nature of the antigen epitope(s) that evoked a biasing of the repertoire during an antibody response.

In this report we have used as an example a previously published collection of >18,000 Ig heavy chain (IgH) sequences from mice immunized with the human complement serine protease C1S (Reddy et al., 2010). Although we have focused exclusively upon an analysis of heavy chain sequences in this example, IgAT is also capable of analyzing human and murine Ig kappa and lambda light chain (IgL) repertoires.

CLONOTYPIC DIVERSITY AS A MEASURE OF RESTRICTION OF THE EXPRESSED REPERTOIRE VERSUS A RANDOM REPERTOIRE

In theory, a diversity of more than 1×10^{15} antibodies can be established from the human and murine Ig germline loci, respectively (Schroeder, 2006). However, several antigen-independent and antigen-dependent mechanisms restrict the expressed antibody repertoires to probably less than 1% of the theoretically available diversity. Current theory holds that during B cell development in



the bone marrow, restrictions are required to avoid the production of harmful or unnecessary antibodies while focusing on potentially protective antibodies. Current data obtained from the deep sequencing of human and mouse IgH repertoires suggests that primary antibody repertoires, while highly diverse, are nonetheless constrained by genetic mechanisms imposed during antigen-independent B cell development (Arnaout et al., 2011; Glanville et al., 2011; Ippolito et al., 2012). A second shift imposed upon the antibody repertoire occurs during the response to antigen. As an indirect measure of divergence from a totally random repertoire, IgAT calculates the clonotypic diversity (clonotypes per functional sequences) and also the sequence diversity (unique sequences per functional sequences). In the example given here, the clonotypic diversity of the IgH chain repertoire after immunization against C1S amounts to 36.6%. In previous studies, we found clonotypic diversities ranging from 27% in extremely immature IgG repertoires from preterm neonates or 30% in IgE transcripts from allergic children up to 81% in peripheral blood IgM repertoires (Zemlin et al., 2007; Kerzel et al., 2010). Although the clonotypic

are stacked in the order of their hydrophobicity according to a normalized Kyte-Doolittle Index (Eisenberg, 1984). Charged amino acid residues are at the bottom, and hydrophobic amino acid residues at the top of each bar as presented previously (Zemlin et al., 2003). **(C)** The Kabat-Wu variability for each position in the CDR-3 region (the higher the score the more variable the position). **(D)** Overall amino acid frequencies within the CDR-3 loop (positions 107–114).

diversity and sequence diversity are essential descriptors of a given sequence collection, the absolute values should only be compared between sequence collections that were obtained with the same method, because (a) increasing rounds of PCR increase the risk of overamplification of a non-representative set of sequences and (b) degenerate V primers may not recognize all V genes with equivalent affinity, leading to a possible underestimation of the true repertoire diversity. Thus, a low clonal diversity might reflect a focusing of the repertoire during oligoclonal or even monoclonal B cell proliferations, but could also be caused by a low number of PCR targets or by suboptimal PCR conditions. Ademokun et al. (2011) have suggested that the reduced clonal diversity observed in peripheral blood IgM, IgG, and IgA repertoires in the elderly might reflect a weaker response to vaccines when compared to young individuals (Ademokun et al., 2011). Moreover, changes of the clonal diversity of the antibody response can be studied longitudinally to characterize the maturation of the antibody repertoire during ontogeny (Zemlin et al., 2007; Kerzel et al., 2010).

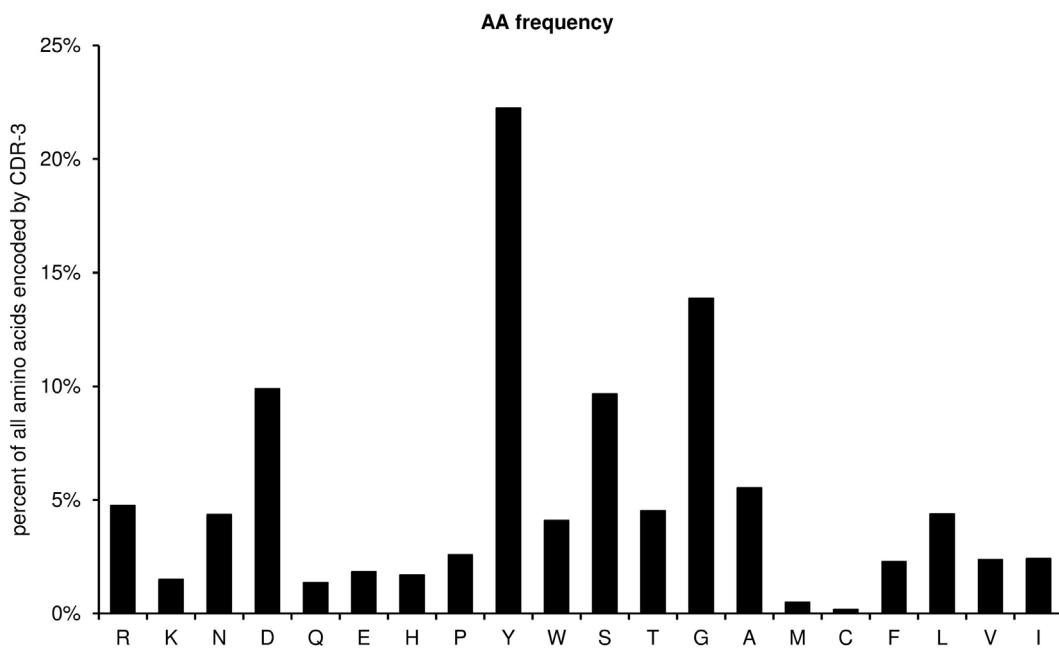


FIGURE 7 | Amino acid frequencies of the CDR-H3 loop for all unique sequences (positions 107–114).

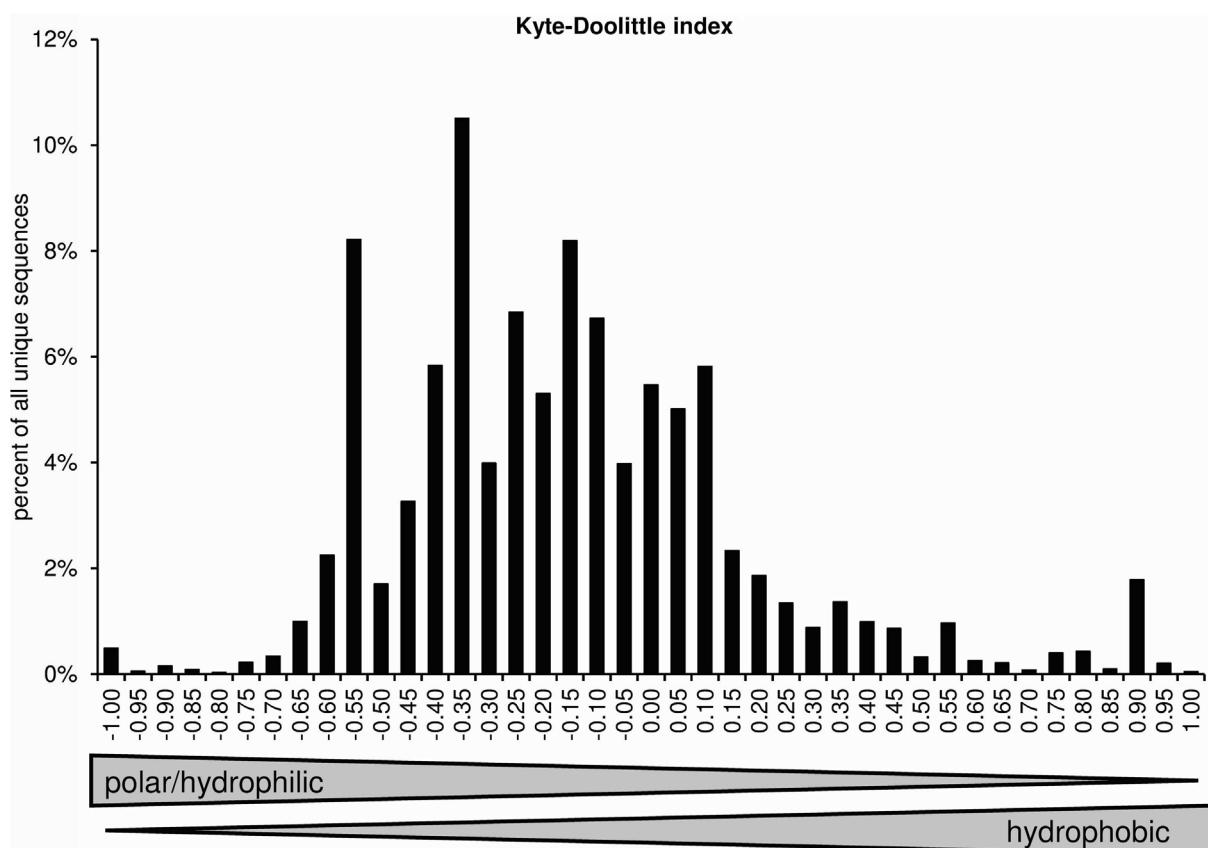


FIGURE 8 | Distribution of average CDR-H3 loop hydrophobicities according to a normalized Kyte–Doolittle scale (positions 107–114; Eisenberg, 1984).

IgAT HELPS IDENTIFYING BIASES IN V, D_H, AND J GENE UTILIZATION THAT CAN INDICATE SUPERANTIGEN-DRIVEN SELECTION OR FREQUENT V_H GENE REPLACEMENT

IgAT summarizes the frequency of V and D_H gene families and individual V, D_H, and J genes. The V_H and V_L gene segments encode for four of the six complementarity determining regions and can thus have great influence on the recognition of classical antigens or superantigens. One reason for contradictory results regarding V gene utilization is the observation that southern blot probes or oligonucleotide primers may not have equal affinity to all V_H gene segments, in particular when somatic mutations affect the primer binding site. To overcome this limitation, Vale et al. (2012)

have suggested a novel technique for a less biased analysis of V_H gene usage. A true predominance of one V gene family or V gene segment can arise from the positive selection of the repertoire for a particular classical antigen or by a superantigen (Zouali, 1995) and has also been described in Ig transcripts of B cell neoplasias (Sasso et al., 1989; Coker et al., 2005; Steininger et al., 2012). The use of individual V_H genes can depend on the position of the V_H gene segments in the germline as well as on epigenetic influences and the multidimensional genomic architecture of the locus (Feeney, 2011). The Ig transcripts studied here as an example were highly polarized toward utilization of the VH1-4 gene segment. Previous analyses demonstrated that this bias reflected the preferred expansion of plasma cells that produced antibodies directed against the antigen used for immunization, C1S (Reddy et al., 2010). Studying V gene expression can also indicate “gaps” in the V gene repertoire that can be a putative cause for increased susceptibility to particular infections such as *Haemophilus influenzae* type B (Feeney et al., 1996). Interestingly, V gene utilization of the Ig heavy and light chains can shift during V gene replacement because only an upstream V gene can replace a downstream V gene (Radic and Zouali, 1996; Zhang et al., 2003). To give a visual impression of a potential 5' shift of V gene usage, IgAT displays V gene segment usage according to each segment's unique position in the germline.

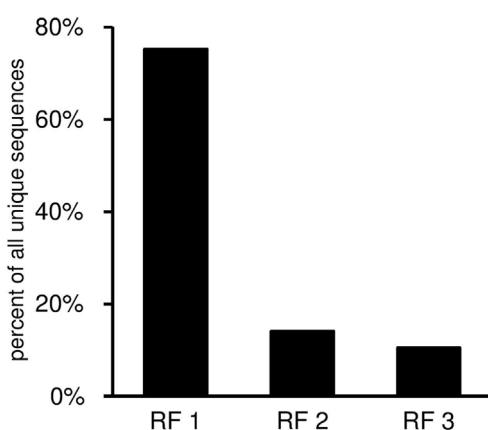


FIGURE 9 | Reading frame utilization given as percent of all unique sequences with identifiable D_H gene segment. The D_H reading frames are defined according to the nomenclature of Ichihara et al. (1989).

BIASES IN AMINO ACID FREQUENCIES AND AVERAGE HYDROPHOBICITY OF CDR-H3 CALCULATED BY IgAT REVEAL RESTRICTIONS WITH POTENTIAL RELEVANCE FOR ANTIGEN RECOGNITION

In the example presented here, IgAT calculated a slightly hydrophilic average hydrophobicity according to a normalized Kyte–Doolittle Hydrophobicity scale for the CDR-H3 region, which is representative for a typical murine primary antibody

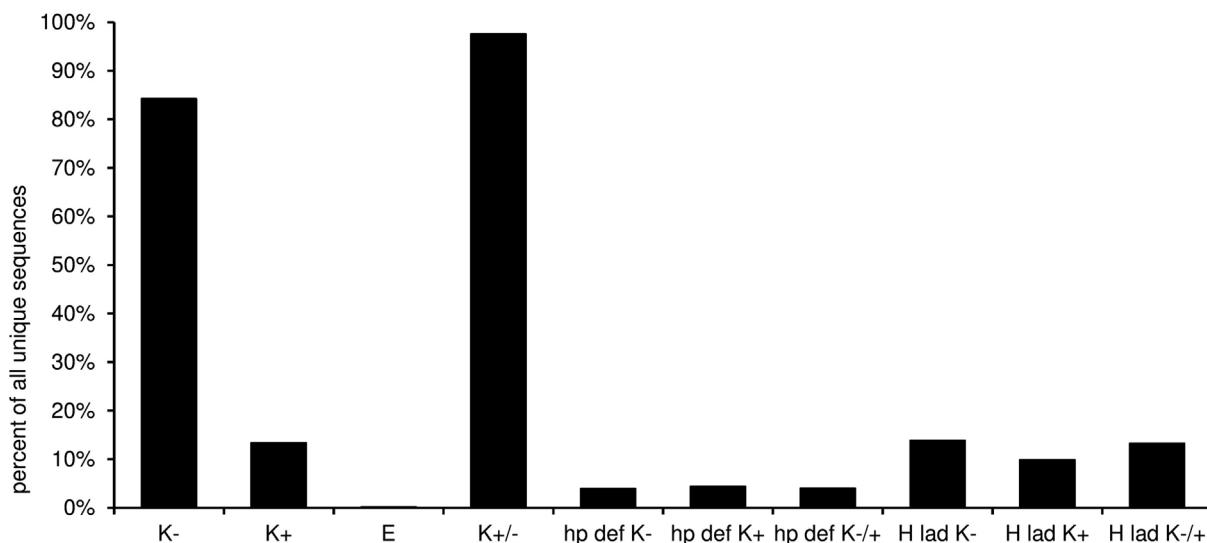


FIGURE 10 | Predicted structural features of the CDR-3 according to the “H3-rules” by Shirai et al. (1999). (K-, kinked base; K+, extra kinked base; K-/, kinked or extra kinked base; E, extended base; hp def K-, deformed hairpin in sequences with kinked base; hp def K+, deformed hairpin in sequences with extra kinked base; hp def K-/+,

deformed hairpin in sequences with kinked and extra kinked base; H lad K-, intact hydrogen bond ladder in sequences with kinked base; H lad K+, intact hydrogen bond ladder in sequences with extra kinked base; H lad K-/, intact hydrogen bond ladder in sequences with kinked and extra kinked base).

repertoire (Zemlin et al., 2003). The hydrophobicity profile of the CDR-H3 region in mice has been shown to be crucial for conservation of global features of a normal antibody repertoire, for generation of normal B cell differentiation, and for the maintenance of normal adaptive immunity to model antigens and pathogens (Ippolito et al., 2006). For example, the position of positively charged amino acids correlates with the specificity against (negatively charged) double strand-DNA in pathogenic autoantibodies (Krishnan et al., 1996) and triplets of hydrophobic amino acids within the CDR-H3 have been implicated with disturbed B cell repertoire formation during a porcine viral infection (Butler et al., 2008). CDR-H3 hydrophobicity is mainly regulated by D_H gene reading frame utilization (reviewed in: Schroeder et al., 2010). The D_H gene segments frequently encode for the core of the CDR-H3, which prototypically lies at the center of the classical antigen-binding site and which therefore can make direct contact with antigen and principally determines Ig specificity (Kabat and Wu, 1991; Padlan, 1994; Xu and Davis, 2000; Collis et al., 2003). Unlike the V and J genes of IgH and IgL loci, the D_H genes are unique in their potential to be used in three forward and three reverse reading frames. The D_H reading frames are characterized by differing hydrophobicity signatures: the first forward reading frame predominantly encodes for hydrophilic amino acids, such as tyrosine, glycine, and serine, and is the most frequent across evolution among jawed vertebrate species, while the hydrophobic second and the often non-functional third reading frame are significantly under-represented (Gu et al., 1991).

Shifts in reading frame usage can be identified by IgAT and may indicate a selective bias regarding the hydrophobicity profile of the antigen-binding site. Moreover, the overall amino acid frequencies of CDR-H3 regions and the frequency of each amino acid per position in CDR-H3 sequences of identical length are presented in bar diagrams by IgAT to characterize a given collection of Ig transcripts and to compare collections that were generated under differing selective pressure.

IgAT ANALYZES THE LENGTH OF CDR-H3 AND ITS COMPONENTS AND CALCULATES PREDICTIONS FOR STRUCTURAL PROPERTIES OF CDR-H3

The CDR-H3 loop can assume an almost unlimited diversity of differing three dimensional shapes which are grouped into canonical structures (Morea et al., 1998). In general, a CDR-H3 region of more than 14 amino acids protrudes into the solvent, while shorter CDR-H3 regions form an antigen-binding groove together with the other CDRs (Ramsland et al., 2001). The three dimensional structure of the antigen-binding site is of great significance for antigen recognition. For example, antibodies directed against virus antigens contain longer CDR-H3 regions on average than antibodies directed against haptens (Collis et al., 2003). Crystallization of antibodies has allowed identifying rules for the prediction of several important structural properties of the H3 loop and of the H3-hairpin based on the deduced amino acid sequence (Shirai et al., 1996, 1999; Kuroda et al., 2008). IgAT applies Shirai's "H3-rules" to predict a kinked, extra kinked or extended shape for the H3 base. The category of the H3 base is mainly determined by the 5' nt of CDR-H3 which are often encoded by the J_H gene. Moreover, for a subset of sequences, the Shirai rules allow the

prediction whether the H3 loop can establish an intact hydrogen bond ladder or a deformed hairpin. In previous studies, we found that intact hydrogen bond ladders were significantly more frequent in IgG heavy chains from preterm neonates than from adults (Zemlin et al., 2007). In both mouse and man, one reason for reduced CDR-H3 length during fetal development is a reduction in the average number of non-templated N-nucleotide additions (Schroeder et al., 1987). Thus, the structural diversity of the H3 loop is heavily restricted during early ontogeny, potentially contributing to the low affinity and poly-reactivity that characterizes the cord blood antibody repertoire.

Besides elucidating the ontogeny of antibody repertoires, the deconstruction of CDR-H3 components provided by IgAT can also give insights into the selective mechanisms during antigen responses. For example, Dorner et al. (1998a) have found that CDR-H3s are generally shorter in non-functional than in functional Ig transcripts and Rosner et al. (2001) observed that mutated Ig transcripts contain shorter CDR-H3s than non-mutated Ig transcripts.

Moreover, IgH receptor editing by the mechanism of V_H replacement result in increased CDR-H3 length due to retention of a portion of the 3' end of the original V_H segment (Zhang et al., 2003). IgAT identifies these " V_H footprints" which tend to accumulate within the V_H - D_H junction during V_H replacement and which typically encode for highly charged amino acids (R, E, and D) at the 5' end of CDR-H3 (Zhang et al., 2003). V_H replacement seems to occur more frequently in autoimmunity (Dorner et al., 1998b).

THE NATURE AND DISTRIBUTION OF SOMATIC MUTATIONS INDICATES ANTIGEN-DRIVEN SELECTION

An enrichment of replacement mutations within the CDRs compared to the FRs is indicative of antigen selection (Berek et al., 1985; Chang and Casali, 1994; Rajewsky, 1996; Lossos et al., 2000). IgAT uses the algorithms created by Chang and Casali (1994), and by Lossos et al. (2000), to identify sequences reflective of antigen-driven selection. In the example given here, 6.5% of the sequences were antigen-selected. This relatively low percentage is plausible since in this experiment, the bone marrow plasma cells were harvested 1 week after immunization, thus before it could be expected that the cells would have undergone excessive class switch recombination and affinity-driven maturation. In previous studies we found that the percentage of antigen-selected transcripts in humans ranged from 9% (IgM) to 29% (IgE) in peripheral blood (Kerzel et al., 2010) and in mice from 0.6% (IgM) to 15% (IgE) in splenic B cells (Rogosch et al., 2010). With this analysis, IgAT quantitatively visualizes the extent to which antigen-mediated selection has impinged upon the B cell repertoire during the course of an immune response.

IN CONJUNCTION WITH IMGT/HIGHV-QUEST, IgAT SIGNIFICANTLY ACCELERATES THE CHARACTERIZATION OF LARGE COLLECTIONS OF Ig TRANSCRIPTS

Fifteen years ago, a researcher needed ~1 h to assign V_H -, D_H -, and J_H -gene segments, N- and P-nucleotides, and somatic mutations to one single Ig heavy chain gene transcript (personal observation). Today, using the freely available IMGT/HighV-QUEST software

and the immunoglobulin gene analysis tool, IgAT, which we present here, it is possible to perform much more detailed analyses on $>10^5$ sequences within hours and $>10^6$ sequences within one day. This comprises only a few minutes of work for the researcher while the remaining time is spent by automated data transfer and analyses. The sequence set used in this report consists of $\sim 18,000$ functional sequences. Results from IMGT/HighV-QUEST were received after ~ 2 h. The calculation time of IgAT depends on the hardware and software configuration of the computer. For example, the analysis takes merely 20 min on an Intel® Pentium® 4 (3 GHz) and 4 GB memory machine running Windows XP (32-bit) and Excel 2010 (32-bit) and 15 min on a AMD® Athlon® 4850e (2.5 GHz) and 4 GB memory machine running Windows 7 (64-bit) and Excel 2010 (32-bit).

REFERENCES

- Ademokun, A., Wu, Y. C., Martin, V., Mitra, R., Sack, U., Baxendale, H., Kipling, D., and Dunn-Walters, D. K. (2011). Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* 10, 922–930.
- Alamyar, E., Giudicelli, V., Li, S., Duroux, P., and Lefranc, M. P. (2012). IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Res.* 8, 26.
- Arnaout, R., Lee, W., Cahill, P., Honan, T., Sparrow, T., Weiand, M., Nusbaum, C., Rajewsky, K., and Koralov, S. B. (2011). High-resolution description of antibody heavy-chain repertoires in humans. *PLoS ONE* 6, e22365. doi:10.1371/journal.pone.0022365
- Benichou, G., Yamada, Y., Yun, S. H., Lin, C., Fray, M., and Tocco, G. (2011). Immune recognition and rejection of allogeneic skin grafts. *Immunotherapy* 3, 757–770.
- Berek, C., Griffiths, G. M., and Milstein, C. (1985). Molecular events during maturation of the immune response to oxazolone. *Nature* 316, 412–418.
- Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., Jones, C. D., Simen, B. B., Hanczaruk, B., Nguyen, K. D., Nadeau, K. C., Egholm, M., Miklos, D. B., Zehnder, J. L., and Fire, A. Z. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.* 1, 12ra23.
- Brezsinschek, H. P., Foster, S. J., Brezsinschek, R. I., Dorner, T., Domiat-Saad, R., and Lipsky, P. E. (1997). Analysis of the human VH gene repertoire. Differential effects of selection and somatic hypermutation on human peripheral CD5(+)/IgM+ and CD5(-)/IgM+ B cells. *J. Clin. Invest.* 99, 2488–2501.
- Brochet, X., Lefranc, M. P., and Giudicelli, V. (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 36, W503–W508.
- Butler, J. E., Wertz, N., Weber, P., and Lager, K. M. (2008). Porcine reproductive and respiratory syndrome virus subverts repertoire development by proliferation of germline-encoded B cells of all isotypes bearing hydrophobic heavy chain CDR3. *J. Immunol.* 180, 2347–2356.
- Chang, B., and Casali, P. (1994). The CDR1 sequences of a major proportion of human germline Ig VH genes are inherently susceptible to amino acid replacement. *Immunol. Today* 15, 367–373.
- Coker, H. A., Harries, H. E., Banfield, G. K., Carr, V. A., Durham, S. R., Chevretton, E., Hobby, P., Sutton, B. J., and Gould, H. J. (2005). Biased use of VH5 IgE-positive B cells in the nasal mucosa in allergic rhinitis. *J. Allergy Clin. Immunol.* 116, 445–452.
- Collis, A. V., Brouwer, A. P., and Martin, A. C. (2003). Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *J. Mol. Biol.* 325, 337–354.
- Cuisinier, A. M., Gauthier, L., Boublé, L., Fougereau, M., and Tonnelée, C. (1993). Mechanisms that generate human immunoglobulin diversity operate from the 8th week of gestation in fetal liver. *Eur. J. Immunol.* 23, 110–118.
- Dahlke, I., Nott, D. J., Ruhno, J., Sewell, W. A., and Collins, A. M. (2006). Antigen selection in the IgE response of allergic and nonallergic individuals. *J. Allergy Clin. Immunol.* 117, 1477–1483.
- Dorner, T., Brezsinschek, H. P., Foster, S. J., Brezsinschek, R. I., Farner, N. L., and Lipsky, P. E. (1998a). Delinement of selective influences shaping the mutated expressed human Ig heavy chain repertoire. *J. Immunol.* 160, 2831–2841.
- Dorner, T., Foster, S. J., Farner, N. L., and Lipsky, P. E. (1998b). Immunoglobulin kappa chain receptor editing in systemic lupus erythematosus. *J. Clin. Invest.* 102, 688–694.
- Dorner, T., and Lipsky, P. E. (2005). Molecular basis of immunoglobulin variable region gene usage in systemic autoimmunity. *Clin. Exp. Med.* 4, 159–169.
- Eisenberg, D. (1984). Three-dimensional structure of membrane and surface proteins. *Annu. Rev. Biochem.* 53, 595–623.
- Feeney, A. J. (2011). Epigenetic regulation of antigen receptor gene rearrangement. *Curr. Opin. Immunol.* 23, 171–177.
- Feeney, A. J., Atkinson, M. J., Cowan, M. J., Escuro, G., and Lugo, G. (1996). A defective Vkappa A2 allele in Navajos which may play a role in increased susceptibility to *Haemophilus influenzae* type b disease. *J. Clin. Invest.* 97, 2277–2282.
- Frolich, D., Giesecke, C., Mei, H. E., Reiter, K., Daridon, C., Lipsky, P. E., and Dorner, T. (2010). Secondary immunization generates clonally related antigen-specific plasma cells and memory B cells. *J. Immunol.* 185, 3103–3110.
- Giudicelli, V., Brochet, X., and Lefranc, M. P. (2011). IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb. Protoc.* 2011, 695–715.
- Glanville, J., Kuo, T. C., Von Budingen, H. C., Guey, L., Berka, J., Sundar, P. D., Huerta, G., Mehta, G. R., Oksenberg, J. R., Hauser, S. L., Cox, D. R., Rajpal, A., and Pons, J. (2011). Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20066–20071.
- Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G. R., Ni, I., Mei, L., Sundar, P. D., Day, G. M., Cox, D., Rajpal, A., and Pons, J. (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20216–20221.
- Gu, H., Kitamura, D., and Rajewsky, K. (1991). B cell development regulated by gene rearrangement: arrest of maturation by membrane-bound D mu protein and selection of DH element reading frames. *Cell* 65, 47–54.
- Hershberg, U., Uduman, M., Shlomchik, M. J., and Kleinstein, S. H. (2008). Improved methods for detecting selection by mutation analysis of IgV region sequences. *Int. Immunol.* 20, 683–694.
- Ichihara, Y., Hayashida, H., Miyazawa, S., and Kurosawa, Y. (1989). Only DFL16, DSP2, and DQ52 gene families exist in mouse immunoglobulin heavy chain diversity gene loci, of which DFL16 and DSP2 originate from the same primordial DH gene. *Eur. J. Immunol.* 19, 1849–1854.

- Ippolito, G. C., Hon Hoi, K., Reddy, S. T., Carroll, S. M., Ge, X., Rogosch, T., Zemlin, M., Shultz, L. D., Ellington, A. D., Vandenberg, C. L., and Georgiou, G. (2012). Antibody repertoires in humanized NOD-scid-IL2Rg-null mice and human B cells reveals human-like diversification and tolerance checkpoints in the mouse. *PLoS ONE*. 7, e35497. doi: 10.1371/journal.pone.0035497
- Ippolito, G. C., Schelonka, R. L., Zemlin, M., Ivanov, Ii, Kobayashi, R., Zemlin, C., Gartland, G. L., Nitschke, L., Pelkonen, J., Fujihashi, K., Rajewsky, K., and Schroeder, H. W. Jr. (2006). Forced usage of positively charged amino acids in immunoglobulin CDR-H3 impairs B cell development and antibody production. *J. Exp. Med.* 203, 1567–1578.
- Jiang, N., Weinstein, J. A., Penland, L., White, R. A. III, Fisher, D. S., and Quake, S. R. (2011). Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc. Natl. Acad. Sci. U.S.A.* 108, 5348–5353.
- Johnson, G., and Wu, T. T. (2000). Kabat database and its applications: 30 years after the first variability plot. *Nucleic Acids Res.* 28, 214–218.
- Kabat, E. A., and Wu, T. T. (1991). Identical V region amino acid sequences and segments of sequences in antibodies of different specificities. Relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. *J. Immunol.* 147, 1709–1719.
- Kalinina, O., Doyle-Cooper, C. M., Miksanek, J., Meng, W., Prak, E. L., and Weigert, M. G. (2011). Alternative mechanisms of receptor editing in autoreactive B cells. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7125–7130.
- Kerzel, S., Rogosch, T., Struecker, B., Maier, R. F., and Zemlin, M. (2010). IgE transcripts in the circulation of allergic children reflect a classical antigen-driven B cell response and not a superantigen-like activation. *J. Immunol.* 185, 2253–2260.
- Kolar, G. R., Yokota, T., Rossi, M. I., Nath, S. K., and Capra, J. D. (2004). Human fetal, cord blood, and adult lymphocyte progenitors have similar potential for generating B cells with a diverse immunoglobulin repertoire. *Blood* 104, 2981–2987.
- Krishnan, M. R., Jou, N. T., and Marion, T. N. (1996). Correlation between the amino acid position of arginine in VH-CDR3 and specificity for native DNA among autoimmune antibodies. *J. Immunol.* 157, 2430–2439.
- Kuroda, D., Shirai, H., Kobori, M., and Nakamura, H. (2008). Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins* 73, 608–620.
- Kurosaki, T., Shinohara, H., and Baba, Y. (2010). B cell signaling and fate decision. *Annu. Rev. Immunol.* 28, 21–55.
- Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* 157, 105–132.
- Lefranc, M. P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G., and Duroux, P. (2009). IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* 37, D1006–D1012.
- Logan, A. C., Gao, H., Wang, C., Sahaf, B., Jones, C. D., Marshall, E. L., Buno, I., Armstrong, R., Fire, A. Z., Weinberg, K. I., Mindrinos, M., Zehnder, J. L., Boyd, S. D., Xiao, W., Davis, R. W., and Miklos, D. B. (2011). High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc. Natl. Acad. Sci. U.S.A.* 108, 21194–21199.
- Lossos, I. S., Tibshirani, R., Narasimhan, B., and Levy, R. (2000). The inference of antigen selection on Ig genes. *J. Immunol.* 165, 5122–5126.
- Morea, V., Tramontano, A., Rustici, M., Chothia, C., and Lesk, A. M. (1998). Conformations of the third hyper-variable region in the VH domain of immunoglobulins. *J. Mol. Biol.* 275, 269–294.
- Morris, G. P., and Allen, P. M. (2012). How the TCR balances sensitivity and specificity for the recognition of self and pathogens. *Nat. Immunol.* 13, 121–128.
- Padlan, E. A. (1994). Anatomy of the antibody molecule. *Mol. Immunol.* 31, 169–217.
- Prabakaran, P., Chen, W., Singarayan, M. G., Stewart, C. C., Streaker, E., Feng, Y., and Dimitrov, D. S. (2012). Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. *Immunogenetics* 64, 337–350.
- Radic, M. Z., and Zouali, M. (1996). Receptor editing, immune diversification, and self-tolerance. *Immunity* 5, 505–511.
- Rajewsky, K. (1996). Clonal selection and learning in the antibody system. *Nature* 381, 751–758.
- Ramsland, P. A., Kaushik, A., Marchalonis, J. J., and Edmundson, A. B. (2001). Incorporation of long CDR3s into V domains: implications for the structural evolution of the antibody-combining site. *Exp. Clin. Immunogenet.* 18, 176–198.
- Reddy, S. T., Ge, X., Miklos, A. E., Hughes, R. A., Kang, S. H., Hoi, K. H., Chrysostomou, C., Hunnicke-Smith, S. P., Iverson, B. L., Tucker, P. W., Ellington, A. D., and Georgiou, G. (2010). Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat. Biotechnol.* 28, 965–969.
- Richl, P., Stern, U., Lipsky, P. E., and Girschick, H. J. (2008). The lambda gene immunoglobulin repertoire of human neonatal B cells. *Mol. Immunol.* 45, 320–327.
- Rogosch, T., Kerzel, S., Sikula, L., Gentil, K., Liebetruth, M., Schlingmann, K. P., Maier, R. F., and Zemlin, M. (2010). Plasma cells and non-plasma B cells express differing IgE repertoires in allergic sensitization. *J. Immunol.* 184, 4947–4954.
- Rosner, K., Winter, D. B., Tarone, R. E., Skovgaard, G. L., Bohr, V. A., and Gearhart, P. J. (2001). Third complementarity-determining region of mutated VH immunoglobulin genes contains shorter V, D, J, P, and N components than non-mutated genes. *Immunology* 103, 179–187.
- Sasso, E. H., Silverman, G. J., and Mannik, M. (1989). Human IgM molecules that bind staphylococcal protein A contain VHIII H chains. *J. Immunol.* 142, 2778–2783.
- Schelonka, R. L., Tanner, J., Zhuang, Y., Gartland, G. L., Zemlin, M., and Schroeder, H. W. Jr. (2007). Categorical selection of the antibody repertoire in splenic B cells. *Eur. J. Immunol.* 37, 1010–1021.
- Schroeder, H. W. Jr. (2006). Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev. Comp. Immunol.* 30, 119–135.
- Schroeder, H. W. Jr., and Cavacini, L. (2010). Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.* 125, S41–S52.
- Schroeder, H. W. Jr., Hillson, J. L., and Perlmutter, R. M. (1987). Early restriction of the human antibody repertoire. *Science* 238, 791–793.
- Schroeder, H. W. Jr., Zemlin, M., Khass, M., Nguyen, H. H., and Schelonka, R. L. (2010). Genetic control of DH reading frame and its effect on B-cell development and antigen-specific antibody production. *Crit. Rev. Immunol.* 30, 327–344.
- Schroeder, H. W. Jr., Zhang, L., and Philips, J. B. III. (2001). Slow, programmed maturation of the immunoglobulin HCDR3 repertoire during the third trimester of fetal life. *Blood* 98, 2745–2751.
- Shannon, C. E. (1997). The mathematical theory of communication, 1963. *MD Comput.* 14, 306–317.
- Shirai, H., Kidera, A., and Nakamura, H. (1996). Structural classification of CDR-H3 in antibodies. *FEBS Lett.* 399, 1–8.
- Shirai, H., Kidera, A., and Nakamura, H. (1999). H3-rules: identification of CDR-H3 structures in antibodies. *FEBS Lett.* 455, 188–197.
- Snow, R. E., Chapman, C. J., Holgate, S. T., and Stevenson, F. K. (1998). Clonally related IgE and IgG4 transcripts in blood lymphocytes of patients with asthma reveal differing patterns of somatic mutation. *Eur. J. Immunol.* 28, 3354–3361.
- Souto-Carneiro, M. M., Sims, G. P., Girschik, H., Lee, J., and Lipsky, P. E. (2005). Developmental changes in the human heavy chain CDR3. *J. Immunol.* 175, 7425–7436.
- Steininger, C., Widhopf, G. F. II, Ghia, E. M., Morello, C. S., Vanura, K., Sanders, R., Spector, D., Guiney, D., Jager, U., and Kipps, T. J. (2012). Recombinant antibodies encoded byIGHV1-69 react with pUL32, a phosphoprotein of cytomegalovirus and B-cell superantigen. *Blood* 119, 2293–2301.
- Takhar, P., Corrigan, C. J., Smurthwaite, L., O'Connor, B. J., Durham, S. R., Lee, T. H., and Gould, H. J. (2007). Class switch recombination to IgE in the bronchial mucosa of atopic and nonatopic patients with asthma. *J. Allergy Clin. Immunol.* 119, 213–218.
- Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature* 302, 575–581.
- Uduman, M., Yaari, G., Hershberg, U., Stern, J. A., Shlomchik, M. J., and Kleinestein, S. H. (2011). Detecting selection in immunoglobulin sequences. *Nucleic Acids Res.* 39, W499–W504.
- Vale, A. M., Foote, J. B., Granato, A., Zhuang, Y., Pereira, R. M., Lopes, U. G., Bellio, M., Burrows, P. D., Schroeder, H. W. Jr., and Nobrega, A. (2012). A rapid and quantitative method for the evaluation of V gene usage, specificities and the clonal size of B cell repertoires. *J. Immunol. Methods* 376, 143–149.

- Vrolix, K., Fraussen, J., Molenaar, P. C., Losen, M., Somers, V., Stinsissen, P., De Baets, M. H., and Martinez-Martinez, P. (2010). The auto-antigen repertoire in myasthenia gravis. *Autoimmunity* 43, 380–400.
- Wu, X., Zhou, T., Zhu, J., Zhang, B., Georgiev, I., Wang, C., Chen, X., Longo, N. S., Louder, M., McKee, K., O'Dell, S., Perfetto, S., Schmidt, S. D., Shi, W., Wu, L., Yang, Y., Yang, Z., Y., Yang, Z., Zhang, Z., Bonsignori, M., Crump, J. A., Kapiga, S. H., Sam, N. E., Haynes, B. F., Simek, M., Burton, D. R., Koff, W. C., Doria-Rose, N. A., Connors, M., Mullikin, J. C., Nabel, G. J., Roederer, M., Shapiro, L., Kwong, P. D., and Mascola, J. R. (2011). Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 333, 1593–1602.
- Wu, Y. C., Kipling, D., Leong, H. S., Martin, V., Ademokun, A. A., and Dunn-Walters, D. K. (2010). High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* 116, 1070–1078.
- Xu, J. L., and Davis, M. M. (2000). Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* 13, 37–45.
- Zemlin, M., Bauer, K., Hummel, M., Pfeiffer, S., Devers, S., Zemlin, C., Stein, H., and Versmold, H. T. (2001). The diversity of rearranged immunoglobulin heavy chain variable region genes in peripheral blood B cells of preterm infants is restricted by short third complementarity-determining regions but not by limited gene segment usage. *Blood* 97, 1511–1513.
- Zemlin, M., Hoersch, G., Zemlin, C., Pohl-Schickinger, A., Hummel, M., Berek, C., Maier, R. F., and Bauer, K. (2007). The postnatal maturation of the immunoglobulin heavy chain IgG repertoire in human preterm neonates is slower than in term neonates. *J. Immunol.* 178, 1180–1188.
- Zemlin, M., Klinger, M., Link, J., Zemlin, C., Bauer, K., Engler, J. A., Schroeder, H. W. Jr., and Kirkham, P. M. (2003). Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J. Mol. Biol.* 334, 733–749.
- Zhang, Z., Zemlin, M., Wang, Y. H., Munfus, D., Huye, L. E., Findley, H. W., Bridges, S. L., Roth, D. B., Burrows, P. D., and Cooper, M. D. (2003). Contribution of VH gene replacement to the primary B cell repertoire. *Immunity* 19, 21–31.
- Zouali, M. (1995). B-cell superantigens: implications for selection of the human antibody repertoire. *Immunol. Today* 16, 399–405.
- Zuckerman, N. S., Hazanov, H., Barak, M., Edelman, H., Hess, S., Shcolnik, H., Dunn-Walters, D., and Mehr, R. (2010). Somatic hypermutation and antigen-driven selection of B cells are altered in autoimmune diseases. *J. Autoimmun.* 35, 325–335.
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Received: 27 April 2012; accepted: 10 June 2012; published online: 28 June 2012.*
- Citation: Rogosch T, Kerzel S, Hoi KH, Zhang Z, Maier RF, Ippolito GC and Zemlin M (2012) Immunoglobulin Analysis Tool: a novel tool for the analysis of human and mouse heavy and light chain transcripts. Front. Immunol. 3:176. doi: 10.3389/fimmu.2012.00176*
- This article was submitted to Frontiers in B Cell Biology, a specialty of Frontiers in Immunology.*
- Copyright © 2012 Rogosch, Kerzel, Hoi, Zhang, Maier, Ippolito and Zemlin. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.*