High-throughput Experimental and Computational Tools for Exploring Immunity and the Microbiome

by

Eliseo Papa

Submitted to the Harvard/MIT Health Science & Technology Institute
in partial fulfillment of the requirements for the degree of

Doctor in Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

Author . . . . . .
Harvard/MIT Health Science & Technology Institute
April 18, 2012

Certified by . . . . . .
Eric J. Alm
Karl Van Tassel Career Development Associate Professor of Civil and Environmental Engineering and Biological Engineering
Thesis Supervisor

Accepted by . . . . . .
Ram Sasisekharan,PhD
Director, Harvard-MIT Division of Health Sciences and Technology. Edward Hood Taplin Professor of Health Sciences & Technology and Biological Engineering

High-throughput Experimental and Computational Tools for Exploring Immunity and the Microbiome

by

Eliseo Papa

Submitted to the Harvard/MIT Health Science & Technology Institute
on April 18, 2012, in partial fulfillment of the
requirements for the degree of
Doctor in Philosophy

Abstract

Humans live in association with trillions of microbes and yet we know remarkably little about their symbiotic relationship. The role these microorganisms have in humans has been characterized only in the case of few bacteria and much less is understood about the dynamic of this relationship. Lately, the mass sequencing efforts accompanying the Human Microbiome Project have begun to uncover the composition of these different microbial niches, and shed light on some the effects they have on their host.

The immune system largely determines the composition of bacterial populations living in association with humans. It fights off pathogens while allowing specific bacteria to colonize the body. However, immune system and microbiota appear even more intimately connected than previously imagined. Recent evidence shows that interaction with the associated microbiota is necessary for the proper development of the immune response throughout life. The interface with commensal microbes is notoriously difficult to probe experimentally, due to the diversity of its composition, which makes differentiating the individual ramifications of each associated microbe a much harder task. To understand the complex relationship between the human immune system and microbiome, we need methodologies that can simultaneously probe both in a high throughput fashion, as well as analysis tools to cope with the large amount of resulting data.

Herein I present the development of immune mass screening tools capable of comprehensively profiling the antibody-mediated and cell-mediated immune response to microbes. I employ microfludics techniques to describe the response of single immune cells at high-resolution and in a physiologically relevant environment. I also present the application of machine learning to gut microbiome data and demonstrate how it can be used to differentiate between diseased and healthy individuals in an IBD patient cohort and to allows to deal with the complexity of microbial community data.

Moving forward, the goal is to combine these approaches to map how changes in the immune

response affects microbiome composition and vice versa. In turn, characterizing this interplay will contribute to our understanding of how bacteria shape our homeostasis and health, facilitating the prediction of which imbalances may lead to disease.

Thesis Supervisor: Eric J. Alm
Title: Karl Van Tassel Career Development Associate Professor of Civil and Environmental Engineering and Biological Engineering

Acknowledgments

A few words of gratitude in exchange for years of support is beyond doubt an unfair exchange. Nevertheless, I shall attempt my best in thanking all the people without whom I could not have finished my studies and have the chance to write this paragraph in the first place. I probably won't be succesful in describing the extent of my appreciation in a short page and for that I apologize in advance. I am also sorry for those who should rightfully be included here and are not. I can only assure them that the role they have had in my life as a graduate student is definitely not forgotten.

First of all I want to thank my advisor, Eric Alm. Since we have met, Eric has never been less than an enthusiastic supporter of my success. His lab was a positive and stimulating place, a perfect spot to come to after what had been a challenging time for me. Eric did not let logistics or bureacracy get in the way of our research together and his sincere curiosity will continue to be an inspiration for many years to come. Thank you for being a mentor, a training partner and a friend: I will be forever grateful. Thanks are also due to all the members of the Alm Lab - Sonia, Jesse, Sean, Arne, Jonathan, Lawrence, Matt, Mark, Ali, Chris, Ilana and Sarah - for all the stimulating conversation and for immediately making me feel at home. I also want to thank Chris Love, who has been a mentor and a friend since the beginning of my graduate studies. Thanks also to my co-authors, the members of the Love Lab and Ploegh Lab, and my colleagues at HST. It was great to share work and fun with all of you.

There are also a few people who deserve particular recognition for the help they have given me through these last few years. Thank you Ludo for sharing in the frustration of scientific research with humour and lightness of being. You are a very thoughtful and important friend. Thank you Antonio for being a true friend and helping me through the toughest and best decision of my life. Thank you Baldur for your friendship and your mixes – which have been the best medicine for my procrastination – with my apologies as such use is diminutive of their value. Thank you Alessandro, for always being there and forever pushing me to grow as an individual. Thank you Tom, Jim, Kim, Sach, Miriam, Juliette, Cha, Eirini, Bernardo, Eduardo, Gonzalo for sharing some great times while

in Cambridge.

Thank you Popi and Spyros for your help. Since the arrival of Isabella, you have helped us make a demanding situation a manageable one. Without your support I would probably still be working on this thesis.

Grazie Mamma e Papa'. Forse solo ora capisco che sacrificio avete compiuto nel lasciarmi andare, nel lasciarmi diventare me stesso. Avete avuto il coraggio di credere ai vostri principi e avete avuto la forza di lasciarmi crescere lontano da voi. Penso di esserci riuscito – sono soddisfatto della persona che sono diventato e di quello che ho dimostrato di saper fare. Sono ben conscio che mai me la sarei cavata senza l'avervi avuto come genitori, come esempio, come costante aiuto.

Nobody however has helped me get here as much as you did Katerina. A new life began for me in front of that conservatory and it is only thanks to you that this life looks so similar to the one I have always dreamed of. I told you that we would be fine, but it was you who had the courage to leave everything and be brave. You put up with the many long nights, supported me at every step, kept me motivated and much much more. How you summoned so much strength and patience is still beyond me. I feel that a thank you will never be enough.

Isabella. Sei arrivata da poco nella mia vita, eppure il tuo dolcissimo sorriso, la tua risata e le tue prime gattonate gia' non lasciano alcun spazio alla tristezza. Grazie, amore mio: mi hai insegnato l'amore piu' grande. Sei felicita' pura e nei tuoi occhi curiosi io vedo tutta la mia speranza. Spero solo di non deluderti come papa', bimba mia.

Contents

Part I

Introduction

Humans are colonized by trillions of bacteria, the vast majority of which exist in a mutualistic relationship with their host. Despite their overwhelming presence on the body, we know little about our associated microbiota. How is the composition of our microbiome determined? How do bacteria colonize the human body and what determines their ecological success within a specific niche? And perhaps most importantly, what functional roles do they play in health and what triggers are required to render them pathogenic?

Even though significant technical advances have increased our ability to map and quantify bacteria, these question remain unanswered in large part due to the lack of a systems approach with which to simultaneously analyze host-derived pressures, microbiome composition and host responses.

1.1 Effects of the immune system on the microbiome

The number and type of bacteria present at different sites on the body are undoubtedly shaped by the nature and interactions amongst the bacterial species themselves, but their composition is also heavily influenced by the human immune response.

Innate immunity – mediated by phagocytes, NK cells and neutrophils – affects the number and species of microorganisms by triggering inflammation. The innate immune system must tolerate associated bacteria while remaining capable of mounting a response to pathogens. This is largely achieved by reducing the contact between the microbiota and the sensing mechanisms of the immune system. This separation is attained through defensive barriers, such as mucus layers and anti-microbial factors. When these barriers fail and contact does occur, sensing of microbial material by toll-like receptors (TLR) present on the surfaces of innate immune cells triggers an inflammatory response. The subsequent release of pro-inflammatory cytokines is responsible for the activation of adaptive immunity and the initiation of bacterial killing by macrophages and NK-cells . These mechanisms, together with the release of lipid mediators of inflammation and reactive oxygen species, are responsible for the consequent reduction in microbial load.

Adaptive immunity – mediated by B cells and T cells – is responsible for the antigen-specific killing of pathogens. B cells primarily secrete antibodies, which can have direct anti-microbial effect (such as IgA in the gut) or otherwise exert an influence on host immune response (such as anti-neutrophil cytoplasmic antibodies). B cells can be activated by cytokine release or by direct binding with their antigen in a native form (through the antigen-specific B-cell receptor or BCR). T cells, on the other hand, are activated by pathogen-derived, MHC-associated peptides via their antigen-specific receptors (T-cell receptor or TCR). Cell-mediated killing occurs through cytotoxic T cells, which interact with targets directly and release factors capable of disrupting pathogen integrity.

Regulation of adaptive immunity is intrinsically linked with the composition and function of the resident microbiota due to its ability to kill specific pathogens. Throughout the gut, various types of antigen-specific T cells have different abundances, which translate into differences in bacterial composition. Two types of T cells are of particular importance in understanding mucosal immunity: Th17 cells and regulatory T cells (Treg). Th17 cells have a role in clearing pathogens during host defense reactions and in inducing tissue inflammation in autoimmune disease. Tregs have opposing effects and are responsible for maintaining a balanced adaptive immune response and prevent excessive microbial killing.

1.2 Reciprocal influences of the microbiome and the immune system

Despite the large influence the host immune system exerts on the microbiota, the communication between host and microbiota is far from unidirectional. It is gradually becoming apparent that – as much as the host influences the composition of the associated microbiota – the presence of microorganisms is in turn required for the proper development of the immune system and regulation of the immune response. Substantial evidence supports a role for enteric microbiota both in protection against disease as well as in its pathogenesis. Imbalances in the equilibrium between host

and microbiota contribute to the development of obesity [1] [2] [3] [4], autoimmune diseases [5] [6] [7] and allergies [8] [9]. Similarly, it is postulated that bacteria behave as pathogens to their host when this equilibrium is perturbed.

However, a complete picture of how this balance is maintained in health and disrupted in disease remains elusive. Understanding the complex interaction between host and microbiota will not only help explain the role bacteria play in human biology, but will also contribute new elements to our grasp of the immune system and the underlying mechanisms of the host response. The ultimate goal would be to enable the manipulation of bacteria for therapeutic purposes.

Current molecular biology tools are not powerful enough to decipher these complex interactions. Traditional immunological assays (ELISA or FACS) provide an overall picture of the immune system, but lack the necessary single cell resolution and large throughput data analysis. Similarly, available microbiological approaches can only focus on a few of the microbes living in association with the human host, largely due to the inherent difficulties of growing representative colonies in culture.

To understand the mechanism of the interplay between host immune system and microbiota it is first necessary to describe each comprehensively and in detail. Herein I present the development of systems biology tools capable of independently characterizing the immune response and the composition of human microbiota with high resolution and accuracy. These tools can be used to define this delicate interplay, thus enabling the discovery of the mechanism of communication between the host and symbiotic bacteria.

1.3 Outline

I begin by reviewing what is currently known about the relationship between host immune system and microbiota as well as the reciprocal influence they exert on each other.

In Part II I present the tools developed for the study of the host immune system: first, a tool to profile antibody-mediated adaptive immune responses with single cell resolution and second,

a tool capable of observing cell surface receptor dynamics in individual lymphocytes. In Part III, I present the computational tools developed for the analysis of microbiota composition and its impact on host health. Lastly, in Part IV I summarize how the tools described further enable the study of host-microbiome interactions and discuss future directions of this work.

1.4 Background

The human microbiome

The human microbiome is established at birth, where maternal bacteria are responsible for the initial colonization. The composition of these microbiota then evolves based on successive environmental exposures. Despite large individual-to-individual variability, it has been determined that a core microbiome of approximately 1200 species is shared amongst humans. [10] It is estimated that the total number of genes contained within these bacterial species is 150 fold the number of genes in the human genome. The majority of the microbiota that lives in association with humans is contained in the gastrointestinal system. It seems therefore plausible that the greater part of the functionalities provided by these microbes reside here.

Humans and microbiome co-evolution

The relationship between the human and the microbiota is mutualistic, therefore the human host depends on the presence of its associated bacteria to thrive in its environment, while the microbes depend on one of the microenvironments in the host for their survival [11]. This mutualism has led to a process of co-evolution which, in turn, has led to the establishment of a balance between the needs of the host and the survival of the bacteria [12]. Bacteria that live in association with humans appear to provide metabolic and digestive functions which would otherwise not be available to the host. They also appear to provide protection from microbial pathogens and, importantly, modulate immune system homeostasis.

Human and microbiota co-evolution is a continuous process and though it may have reached an equilibrium over time, it is not known how long this state has lasted or how stable it will be over time. In fact, it has been hypothesized that many of the perturbations of this balance we observe today are the manifestation of this evolutionary competition between host and microbes [13] [14]. Temporary evolutionary advantages of microbes over their host may manifest as pathogenic processes. Similarly, flexible mechanisms of immune response may be an evolutionary adaptation of the host to the rapid pace of genetic change in the microbiota [15].

Ecological relationships in the microbiome

It is thought that microbes living in association with the human host are in a relationship spanning the continuum from pathogenic to symbiotic [16]. In a *symbiotic* relationship at least one between microbe or host would benefit from co-existence with no adverse effect on the survival of the other. In the gut, for instance, microbes can work symbiotically with their host, providing protection from other microorganisms and aiding in the breakdown of compounds the host would otherwise be unable to metabolize.

At the other end of the spectrum lie *pathogenic* microbes, which subsist on and potentially harm the host while thriving in the niche it provides. Somewhat in the middle of this continuum there lies a category of microbes termed *commensals* which inhabits the host without posing either obvious harm or benefit to it. Lastly, it has been suggested that a new category of bacteria termed *pathobiont* should be defined [11] [17]. These are normally symbiotic bacteria which, under particular environmental circumstances, are capable of behaving as a pathogen causing harm to their host.

One prototypical example of a pathobiont in humans is *Helicobacter pylori*, which asymptomatically colonizes the gastrointestinal mucosa, without causing any secondary harm [12]. In tandem, *H. pylori* has been demonstrated to be a major cause of gastritis and gastric cancer in humans. Although the reasons for this incongruence are not completely understood, it is thought that only particularly aggressive strains of *H. pylori* can cause disease and do so in the subset of hosts that

have a genetical predisposition. Other bacteria behaving in a similar manner have also been identified in mice (*Helicobacter hepaticus* [17] [18] and enterotoxigenic *B. fragilis* [19]) and in humans (adherent *Escherichia Coli* strains [20]).

The immune system interface with the microbiome

Study of the relationship between microbiota and gut immune system are beginning to elucidate the molecular mechanism by which the immune system is capable of tolerating such an extensive exposure to microbial material and yet remain capable of mounting an immune response to pathogens. The mechanism through which this tolarance occurs has been termed immune *ignorance* [8]. Rather than developing immune *tolerance* through sustained exposure to microbial factors, the mucosal immune system relies on being largely shielded from the commensal bacteria inhabiting the gut. Immune ignorance to the microbiota is achieved by a combination of anti-inflammatory processes, initiated either by the host immune cells or the commensal microbes themselves, and the physical barriers that limit the majority of bacteria to the lumen of the gut and away from the mucosal layer [11]. The lack of an inflammatory process prevents antigen-presenting cells from engulfing commensals and, in combination with physical barriers, minimizes the contact between immune system and bacteria under normal conditions.

Antigen presenting cells, NK cells and regulatory immune cells are responsible for the release of anti-inflammatory mediators, which prevent contact with microbes from eliciting a full immune response. This paradoxical mechanism has most likely evolved to protect the host from the consequences of excessive and prolonged immune responses, including severe epithelial damage and loss of integrity of the gut mucosa. Resident microbes have also evolved strategies to prevent inflammatory states in the immune system, since inflammation would bring about the indiscriminate elimination of microbes in the mucus layer as well as in the lumen.

Mucus is the primary physical barrier that prevents microbes to enter in contact with the epithelial cells of the gut. It is largely composed by the MUC2 protein and it is divided into two layers: an outer layer in contact with the microbiota and the lumen, and an inner layer which is

hard to penetrate for bacteria and almost completely devoid of microbial load in vivo [**?**].

Despite the existence of these innate defenses, the separation between microbiota and immune system is not absolute, since some components of the immune system (eg M cells in peyer's patches and DC cells in the lamina propria) have been shown to sample commensal bacteria under normal circumstances [21]. The implications of this immune recognition however still remain unclear [22].

Microbiome influence on host immune system

Evidence is mounting that microbiota shapes the immune response in fundamental ways. The study of the influence of microbiota on the host immune response has mainly focused on the gut, where it has been demonstrated that bacteria are required for the proper development of mucosa-associated lymphoid tissue. In germ free mice, where the contribution of the microbiota is absent, the gut-associated lymphoid tissue (GALT) does not fully develop [23].

Furthermore, commensal microbes have been shown to directly modulate immune homeostasis, both at the level of gut mucosal immunity and systemic immunity [24] [25]. There are very large numbers of species in the microbiota, ecological competition is tight and, therefore, it is not surprising that microbial species have evolved ways of influencing the host immune system in order to maximize survival. The evolution of such mechanisms is likely an evolutionary strategy on the part of the microbes in order to preserve mutualism [26] [27], however this remains a difficult hypothesis to test.

The impact of commensals on the balance of the host immune system appears to be subtle under normal conditions, but becomes an important determinant of host fitness in situations where the host and its gut are under the influence of environmental stress [2]. The microbial composition and the shape of the immune response mounted can determine the host's fitness during infection or inflammation.

Mechanisms

Commensals favor proper functioning of the immune system in the gut by a variety of mechanisms [28]. Mutualistic bacteria may exert their effect either by i) direct penetration of the mucosal membranes and engagement with epithelial cell receptors [29] or ii) by secreting specific metabolites [30].

When engaging with epithelial cells, mutualistic bacteria contribute to the integrity of the gut protective barriers, by influencing the rate of proliferation of intestinal epithelial cells and inducing the production of cytoprotective proteins. Resident bacteria in the mucus layer also stimulate the release of anti-microbal compounds capable of significantly reducing bacterial load in the lumen [31]. Microbial-associated molecular patterns (MAMPs) are recognized by innate immune receptors which then stimulate the release of anti-microbial peptides (AMPs) and secretory IgA[ˆiga-note] — through BAFF and APRIL transcription factors [32] [33].

Direct interaction between cells of the innate immune system and the bacteria living in the gut primarily involves toll-like receptor (TLR) signaling. Loss of TLR signalling in MyD88 deficient mice causes greater susceptibility to colonic injuries, suggesting a role of TLR signalling in the maintenance of epithelial integrity in the gut [34].

The secretion of microbial factors or metabolites is also responsible for changes in the mucosal immune system. For instance, it has been shown that bacterial ATP has an effect on Th17 cell differentiation [35]. Similarly, bacterial peptidoglycans have been shown to enhance systemic immunity [36] and soluble proteins produced from probiotic bacteria have been shown to regulate the integrity of the intestinal epithelium [37].

Lastly, the balance between regulatory T cells and Th17 cells appears to be another important mechanism by which the microbiota exert their control over mucosa. Tregs are a population of T cells which is abundant in the gut and present also in the absence of microbiota. It is believed that their differentiation is influenced in large part by the gut microbiota, since DNA from specific commensal bacteria has been shown to directly reduce Treg differentiation [38]. Tregs are less

abundant in the small intestine than in the colon. Th17 on the other hand are less abundant in the colon than in the small intestine. It is believed that the different makeup of the microbial communities at these two sites influences the balances between these two pools of cells and consequently, the characteristics of the T cell mediated immune response [18].

Effects

It has been demonstrated that different bacteria modulate the immune system by independent mechanisms. The two most important examples are *Bacteroides fragilis* and segmented filamentous bacteria (SFB). *B. fragilis* can induce Th1 responses in the gut through polysaccharide-A (a surface polysaccharide specific to *B. fragilis*) which in turn stimulates proliferation of regulatory T cells. Thus its main effect is to promote an anti-inflammatory state in the gut [17] [39]. SFB arguably acts in the opposite direction, stimulating the proliferation of Th17 cells in the gut and thus promoting strong effector immune responses [40]. As a result, animals whose microbiota is devoid of SFB have worse resistance to some pathogenic infections [41]. On the other hand, colonization with SFB worsens inflammation in mice models of autoimmunity and can exacerbate disease [11]. Other specific bacteria have been postulated to have protective effects in inflammatory disease and colitis (*F.prauznitsii* [42] and *Clostridium* bacteria) or colitogenic effect (*K.pneumoniae* in SC mice) [24], though more details about the underlying molecular mechanism have yet to be clarified.

However, the effective role of each bacterium is often not so clear cut. The overall composition of the microbial community can change the behaviour of a single microbe, as shown by experiments where SFB alone was not capable of triggering any intestinal inflammation in immunodeficient mice [12]. Similarly, the integrity of the immune system can ultimately determine the effect of each microbe on the host. Even the prototypical anti-inflammatory bacterium *B. fragilis* has been shown to cause sepsis in an immunocompromised host [43].

Importantly, the observation that distinct microbes have different effects on the immune system validates the possibility of using individual microorganism to manipulate the host-microbiota

balance for therapeutic uses. Moreover, it underlines the importance of determining microbiota composition to fully characterize immune response. If the presence or abundance of specific bacteria can result in different overall immune responses to pathogens, it will be important to determine which bacteria live in association with the host in order to predict the success and duration of the response.

Inflammatory bowel disease

Inflammatory bowel diseases (IBD) - ulcerative colitis (UC) and Crohn's disease (CD) - are diseases characterized by the inflammation of the colon and small intestine. The exact cause and etiology of IBD is unknown, and diagnosis is made on the basis of symptoms, particularly the presence of lesions in the lower GI tract. It is considered an autoimmune condition and is usually managed through immunosuppressive therapies.

There is increasing evidence that host-associated microbiota plays a key role in IBD. An inverse correlation between incidence of parasitic gut microbiota, such as worms, and the occurrence of IBD, suggests a protective effect of parasites on host [44]. Furthermore, prebiotics and probiotics also have beneficial effects in patients with IBD, in some cases more so that conventional drug regimens [5] [45]. At the same time, microbiota has also been linked to exacerbated disease pathology, with ulcerations occurring in stretches of the intestines with the highest concentration of bacteria [46]. In certain cases, treatment with antibiotics, and therefore reduction of gut microbes, can diminish IBD symptoms.

Dysbiosis in inflammatory bowel disease

To explain the critical influence of the microbiota in inflammatory bowel disease, the concept of *dysbiosis* has evolved. This is an imbalance in the flora associated with disease, which can be equally brought about by environmental factors, colonization by an invasive pathogen, or T-cell mediated host immune response, as demonstrated in the case of *Salmonella Enterica* [47]. These imbalances have been largely characterized in mouse models as well as humans [7], but are rarely con-

served across individuals. A few scenarios for bacterial-dependent IBD etiology have been postulated including i) an increased prevalence of pro-inflammatory bacteria capable of initiating disease episodes or worsening existing inflammation, ii) a reduction of those taxa responsible for inhibiting inflammatory cascades in intestinal epithelial cells, iii) abnormal adherence of the intestinal microbiota to the epithelial wall, iv) an overall decrease in diversity and v) selection of colitogenic bacteria by intestinal inflammation [**?**]. It is perhaps more likely that all these scenarios simultaneously play a part in the etiology of IBD, though the extent of each contribution's changes depends on the stage of the disease (before or after inflammation), the overall architecture of the commensal microbiota, the presence of other environmental stressors (eg. diet, infections, antibiotics) and predisposing genetic factors in the host (eg. immune deficiencies, NOD2 variants). Since these processes are reciprocally connected, pinpointing the actual trigger causing the onset of disease may be a challenging task. It is indeed probable that the triggering event may be different in each IBD patient and that an integrated ecological approach may be more appropriate in the treatment of IBD.

Dysbiosis and systemic immunity

Dysbiosis has also been implicated in extra-intestinal disease, particularly in allergy and autoimmune diseases. Hence, there is increasing interest in systematically determining how microbiome correlates with changes in systemic immunity. Despite convincing proof that immune response to intestinal microbiota is largely confined to the GALT, there seems to exist a different mechanism by which gut microbes affect the development of systemic immunity in the host. This hypothesis derives from the observation of significant differences in the immune system of germ-free mice when compared to control animals. Translocation of microbially-derived factors from the intestinal mucosa to the lymphatic system could account for this influence [48] [36], but more research is required to obtain a clear picture.

Regardless of the mechanism by which the microbiota causes these systemic changes in the immune response, they provide a plausible mechanism for the correlations observed between mi-

crobial dysbiosis and **autoimmune diseases**. [49] Studies in mice have shown how microbiota can either worsen or ameliorate symptoms of autoimmune disease, though they failed to provide a casuative link. Colonization with microbiota was shown to prevent diabetes in germ-free MyD88-/-NOD mice, while at the same time increasing the severity of autoimmune encephalomyelitis.

Allergic reactions have also been correlated with microbial imbalances as part of the effort to demonstrate the *hygiene hypothesis*[^hygienehypothesis]. Administration of probiotics in children can reduce IgE reactions, albeit temporarily [50]. Another study in mice models of allergy demonstrated the role of microbial LPS in preventing anaphylaxis, and showed that TLR9 stimulation could abolish allergic symptoms [51]. These correlations between microbial imbalances and systemic immunity have prompted research into the impact of the microbiome on the efficacy of **vaccinations**. On one hand, there is interest in determining if routine vaccinations have an impact on the development of microbiota, though this hypothesis seems now largely discredited. [52] On the other hand, given the impact of intestinal microbiota in the development of immune function, it is likely that the intestinal microbiota will significantly affect how individuals respond to vaccine antigens. Some studies have indicated a correlation between bacterial overgrowth in the gut with the efficacy of live-cholera vaccines [53], but it remains difficult to dissect to what extent these effects are due to the microbial differences or other unmapped environmental factor. [54] There is now keen interest in manipulating microbiota to either boost systemic response to vaccines or as an adjuvant and delivery vehicle for new vaccines. Though experimental studies have shown promise, no clinical data has yet been reported. [52]

Part II

Tools for the study of host-associated microbiota

Until recently, our understanding and study of microbial communities and their diversity have been severely limited. The majority of naturally occurring microbes are refractory to laboratory culture, thus biasing the study of these populations towards those members that can be individually propagated.

Novel and less expensive sequencing technologies are now poised to dramatically expand our understanding of the microbial world. Through genome sequencing and functional classification, it is now possible to characterize entire populations of microbes, in effect identifying each individual member [55]. Amplifying and sequencing microbial DNA in a complex community, allows for quantification of the extent of genetic diversity in a specific environment, mapping of the underlying structure of the microbial population, and provides clues into the mechanisms responsible for its maintenance [56][57], the communication amongst its members [58] and finally, its evolution [59].

Following these advances, an increasing number of studies of microbial diversity has been conducted and large amounts of data are beginning to accumulate. However, to infer any biological insight from the large amount of sequencing data generated it will be necessary to implement superior analytical methods capable of isolating important features in the dataset.

Machine learning algorithms represent a promising approach, as they allow for the automatic inference of discriminating characteristics and defining features from large amounts of complex structured data. The work demonstrated here aims to apply and adapt existing machine learning techniques to the analysis of microbial diversity.

Chapter 2

Non-invasive Mapping of the Gastrointestinal Microbiota Identifies Children with Inflammatory Bowel Disease.

Eliseo Papa, Michael Docktor, Christopher Smillie, Sarah Weber, Sarah P. Preheim, Dirk Gevers, Georgia Giannoukos, Dawn Ciulla, Diana Tabbaa, Jay Ingram, David B Schauer, Doyle V Ward, Joshua R Korzenik, Ramnik J Xavier, Athos Bousvaros, Eric J Alm

This chapter is presented as submitted to [PLoS ONE]

Corresponding supplementary material is integrated with the text

Abstract

Pediatric inflammatory bowel disease (IBD) is challenging to diagnose because of the non-specificity of symptoms; an unequivocal diagnosis can only be accomplished using colonoscopy, which clinicians are reluctant to recommend for children. Diagnosis of pediatric IBD is therefore frequently delayed, leading to inappropriate treatment plans and poor outcomes. We investigated the use of 16S rRNA sequencing of fecal samples and new analytical methods to create a sensitive and specific test for IBD. We applied supervised learning in microbial ecology (SLiME) analysis to 16S sequencing data obtained from i) published surveys of microbiota diversity in IBD and ii) fecal samples from 91 children and young adults who were treated in the gastroenterology program of Children's Hospital (Boston, USA). The developed method accurately distinguished control samples from those of patients with IBD; the area under the receiver-operating-characteristic curve (AUC) value was 0.83. The accuracy was maintained among data sets collected by different sampling and sequencing methods. The method identified key taxa associated with disease states and distinguished patients with Crohn's disease from those with ulcerative colitis with reasonable accuracy. The findings were validated using samples from an additional group of 77 patients; the validation test identified patients with IBD with an AUC value of 0.76. Microbiome-based diagnostics can distinguish pediatric patients with IBD from patients with other gastrointestinal conditions. Although this test should not replace endoscopy and histological examination as diagnostic tools, classification based on microbial diversity is an effective complementary technique for IBD detection in pediatric patients.___

2.1 Introduction

Crohn's disease (CD) and ulcerative colitis (UC), collectively termed inflammatory bowel diseases (IBD), are incurable conditions that cause ulceration of the intestinal mucosa. If left untreated, IBD may require repeated surgical intervention to remove affected parts of the gastrointestinal system [60] leading to malabsorption and nutritional complications. [61] Despite its importance, timely diagnosis is difficult because patients often present with non-specific symptoms, [62] and the presence of CD or UC can only be confirmed by colonoscopy.

Diagnosis is particularly challenging in children, for whom presenting symptoms may vary widely and may only consist of subtle extra-intestinal manifestations. [63] This in turn leads to a typical delay in the diagnosis of pediatric IBD, ranging from 4 weeks in severe colitis [64] to 6–7 months in milder disease. [63] Reducing this diagnostic delay is important, since a long period of unmanaged symptoms can significantly impact growth [64] and early treatment is essential to preserving long-term quality of life. [65] Thus a sensitive yet non-invasive detection tool, that could identify patients at high risk for IBD, and therefore warranting endoscopic evaluation, would be a valuable diagnostic aid.

Non-invasive tests for IBD already exist, including antibodies [66], imaging-based screens, [67] [68] and fecal biomarkers extbackslashcite [69]. Specificities for existing methods range from 89% to 95% for either CD or UC, [70] however,these methods are either limited to active disease, poorly sensitive ($\sim$55%), or their outcome can be confounded by diseases other than IBD, [70] limiting their clinical utility. [71] [72]

The design of an accurate test for IBD is challenging, since the precise cause of IBD is unknown. No single genetic, environmental or epidemiological factor alone is diagnostic of IBD. [73] Instead, current evidence about the aetiology of IBD points toward a complex interplay between genetic, environmental, and immunological factors [74] [75] [46] and the intestinal microbiota. [76] [5] [77] Arguing in favour of the involvement of gut microbes in the pathogenesis of IBD, it is known that colonisation with commensal bacteria is required to elicit colitis in mice. [5] [78] Similarly,

in IBD patients it is known that antibiotics can treat CD colitis in the short term [79] and probiotics may prevent relapse of UC. [80] We hypothesized that changes in the intestinal microbiota, whether causative of or responsive to disease, may provide a viable diagnostic of disease status.

Previous microbial diversity studies have found characteristic changes in the composition of the gut flora during IBD that could potentially be used to screen patients with non-specific symptoms. [76] [7] In one of the most comprehensive studies to date, Frank and colleagues [7] mapped microbiota composition in 124 IBD and non-IBD patients by biopsy sampling coupled with 16S rRNA sequencing. Their work showed that patients with a long-standing history of IBD had decreased levels of Firmicutes and increased level of Proteobacteria, when compared to control individuals. While these results firmly established the relationship between GI microbiota and disease status, the overall approach is unsatisfactory as a diagnostic tool because of low sensitivity (31%) and low overall accuracy (51%, as determined from Fig. 3 in [7]).

More recent studies have been able to accurately distinguish CD and healthy individuals on the basis of pyrosequencing data, [81] but the same model was unable to distinguish UC from healthy individuals or to differentiate patients in remission from patients with active disease, raising questions about whether such approaches show clinical potential. Finally, none of these studies examined pediatric cohorts.

Here we demonstrate an approach that is capable of routinely differentiating children with IBD from controls with other gastrointestinal diseases in a case-control study of ninety-one pediatric patients. Our methodology shows high sensitivity and specificity over a range of disease prevalence and it can be used to i) identify key taxa associated with each disease state, ii) discriminate CD and UC and iii) differentiate patients with active disease from those in remission. We confirmed our results by blind validation with an independent cohort of seventy-seven pediatric patients. This method applies next-generation sequencing and robust statistical analysis using machine learning techniques and, significantly, is a test for IBD based on non-invasive fecal sampling.

2.2 Results

Supervised classification distinguishes IBD and non-IBD patients in existing tissue-based studies

The case-control study of Frank and colleagues [7] used an unsupervised clustering approach: principal components analysis (PCA). When the class labels (healthy vs. diseased) are known for a training set of samples, then supervised learning methods are preferred (e.g. support vector machines, random forests, etc.). These algorithms have been widely and successfully applied to many problems in the biomedical sciences [82] and their use in a clinical setting is emerging in the analysis of gene expression data [83] and microbiome data. [84]

We first investigated whether supervised learning could improve the performance of a microbiota-based IBD detection tool, by applying it to the published IBD data set of Frank et al. [7] We employed our Supervised Learning in Microbial Ecology (SLiME) method to classify samples from the existing data set as IBD or control. SLiME is based on Random Forests (RFs), [85] although we achieved similar results using other supervised learning approaches such as bagging, stacking and support vector machines (see Figure 2.1).

Applying our classification algorithm to the existing data set yielded a marked improvement in accuracy when classifying patients into IBD or non-IBD groups, mainly by allowing the model to be tuned for increased sensitivity. Based on repeated ten-fold cross validation, the area under the ROC curve (AUC) — which is a measure of the overall accuracy of the classification algorithm over the range of possible disease incidence [86] — was 0.73 (Figure 2.2). Choosing a cutoff on the curve which gives relatively high sensitivity (ie. 87.6%) yields 47.3% specificity, amply surpassing the accuracy of the clustering method originally employed. [7]

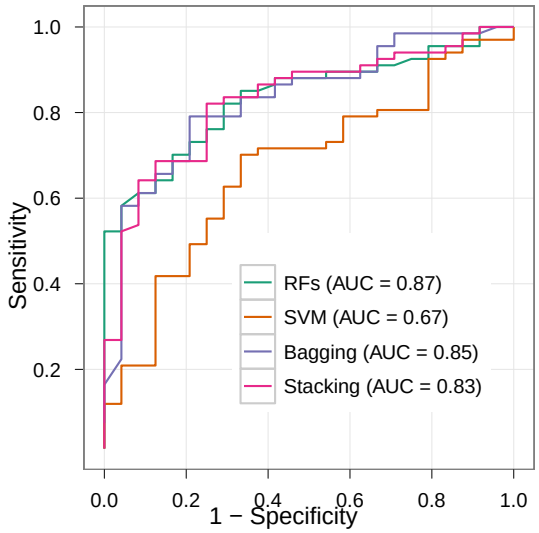

Figure 2.1: Patients are classifiable as IBD and non-IBD with a variety of supervised learning algorithms. ROC curves for SVM, Bagging, Stacking and RFs are shown

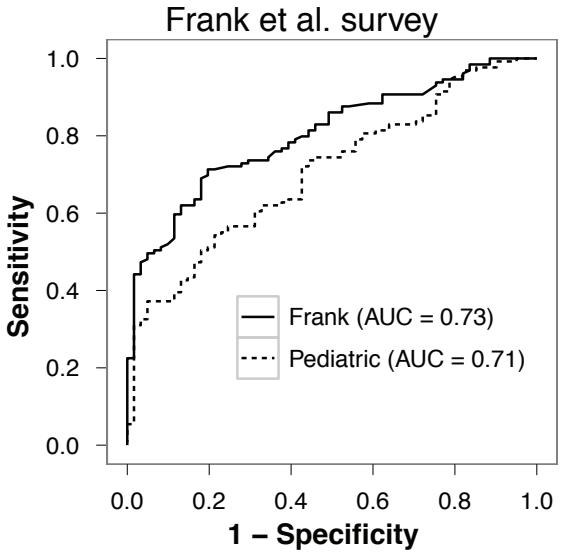

Figure 2.2: Accuracy of disease classification. ROC curve for SLiME classification of active IBD patients vs controls in the pediatric case-control data set.

Supervised classification distinguishes IBD and non-IBD pediatric patients on the basis of stool

Although the results obtained using the existing Frank et al. data set were encouraging, there are several reasons why they might not translate to a clinically useful diagnostic test. First, samples were obtained invasively through surgical tissue resection from adult patients with advanced disease, and may not reflect changes observed in fecal samples from patients with less advanced disease. Second, the control specimens in the Frank et al. study were largely composed of tissue from cancer patients, and thus were not typical of patients investigated for IBD in the pediatric setting. We therefore designed a new case-control study to evaluate whether fecal samples from children not undergoing surgery could be utilized to differentiate between patients with and without IBD.

We selected a group of ninety-one children and young adults receiving care in the gastroenterology program of Children's Hospital (Boston, USA), and obtained fecal samples. Of these children, 23 had Crohn's disease, 43 had ulcerative colitis, one had undefined IBD (colitis with elements of CD and UC) and 24 had non-IBD functional disease (patients with gastrointestinal symptoms but no intestinal inflammation). Demographics of the patient populations are given in Table 2.7 We isolated DNA from the fecal samples and sequenced a portion of the 16S rRNA gene using high throughput 454 pyrosequencing (see Materials & Methods). We then applied SLiME to the resulting microbial compositional data.

Remarkably, performance of our method improved on this data set despite the substitution of mucosal samples with stool samples, yielding a ten-fold cross-validated AUC of 0.83 for distinguishing IBD patients from controls (Figure 2.3). Sensitivity and specificity for the diagnostic test can be obtained by selecting the desired threshold along the curve. For instance, choosing a cut-off on the curve at relatively high sensitivity (Figure 2.3, circle) yields 80.3% sensitivity and 69.7% specificity for the test.

The performance of the same classification algorithm was higher when it was applied to distinguish from controls only those IBD patients with clinically active disease, yielding an AUC of

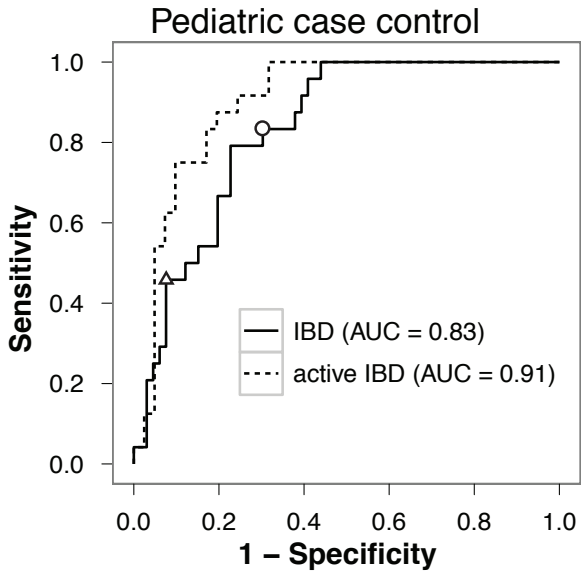

Figure 2.3: SLiME applied to Frank et al. biopsy data set. The black line indicates performance obtained when features were generated by taxonomical binning of the original sequence data (AUC = 0.73); dashed line shows performance when features were selected based on their importance in the pediatric case-control data set and then applied to the Frank et al. study (AUC = 0.71). Two different threshold selections are highlighted: circle, for which SLiME has 80.3% sensitivity and 69.7% specificity; triangle, for which SLiME has 45.8% sensitivity and 92.4% specificity.

0.91 (Figure 1B dashed line). Table 2.1 and Table 2.2 show how the classifier performs amongst the three disease groups (CD,UC and control) at one arbitrary threshold.

Table 2.1: Confusion matrix for the SLiME classification of the pediatric training cohort. Sensitivity 87.6%. Specificity 45.8%. Note this is only one possible cutoff value. Different sensitivity and specificity can be obtained by appropriately tuning the cutoff

	SLiME classification	
Diagnosis	IBD	non-IBD
CD	17	5
UC	40	3
Control	13	11

Table 2.2: Confusion matrix for the SLiME classification of the training cohort on the subset of patient with clinically active disease at the time of sampling. Sensitivity 82.5%. Specificity 75%. Note this is only one possible cutoff value. Different sensitivity and specificity can be obtained by appropriately tuning the cutoff.

	SLiME classification	
Diagnosis	IBD	non-IBD
CD	3	5
UC	30	2
Control	6	18

To test if the chosen sequencing technology altered the classification of patients into controls and IBD samples, we repeated sequencing for 10 of the samples using the Sanger sequencing method. Supervised learning results, however, were independent of the sequencing method employed (Figure 2.4).

We hypothesized that some of the improvement in performance might be due to increased sampling depth if a subset of discriminatory bacteria are present at low abundance. To test this hypothesis, we identified the bacterial taxa most strongly associated with IBD (either positively or negatively), and plotted their abundance. As shown in Figure 2.5, many of the most informative taxa are present at a level of less than 1% per sample – the level at which we would expect to see

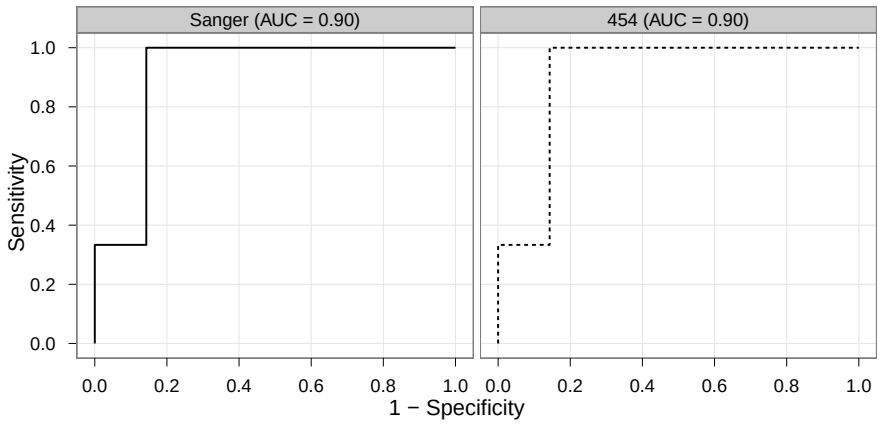

Figure 2.4: Sequencing technology does not significantly influence classification accuracy. ROC curves for active IBD vs. control classification in ten samples where sequencing was repeated by Sanger methods and yielded the same area under the curve.

one count or less at the sequencing depth used in Frank et al. [7] Thus, sequencing depth is an important factor in diagnostic accuracy and may account to a large degree for the lower AUC we obtained in the classification of the Frank et al. data set.

Distinctive taxonomical groups are associated with IBD

We identified a number of bacterial taxa strongly associated with IBD that both confirmed and supplemented previous studies. Figure 2 shows taxa that are significantly associated with either IBD or control patients (q-value < 0.01, Kruskal-Wallis test, FDR adjusted, [87] $E(\pi 0)=0.18$, see Figure 2.6).

Only a few of these taxa show a distribution consistent with an ideal microbial biomarker – a bacterial group whose presence/absence indicates disease phenotype. For example, the Enterobacteriales are indicators of active IBD (ie. patients with clinically active disease and not in remission), while Rikenellaceae and Porphyromonadaceae are generally found within the control group. By contrast, most of the discriminatory groups in Figure 2.7 are more or less abundant in IBD patients, but not exclusive to one population (e.g., the Butyricicoccus and Subdoligranum

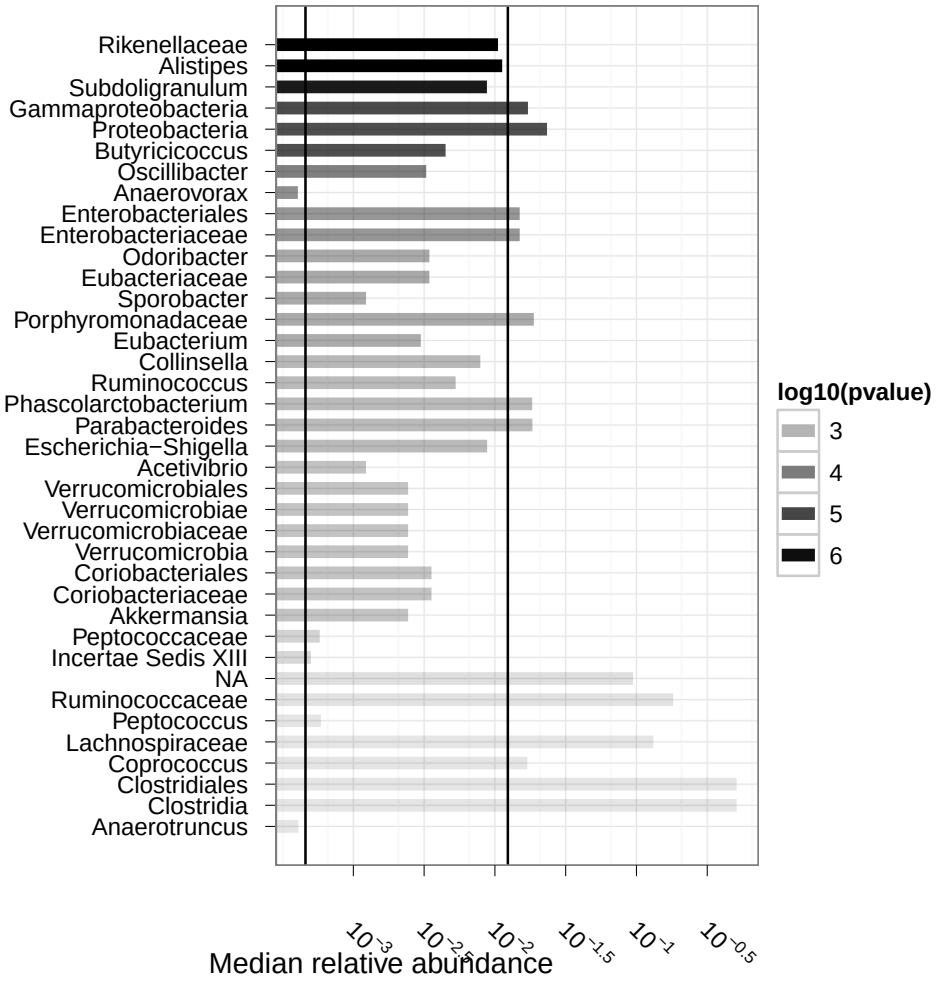

Figure 2.5: Relative abundance of each discriminatory feature compared to the sequencing depth of other IBD microbiota surveys. Two vertical lines indicate the minimum detectable abundance in the Frank et al. study (right) and the Willing et al. study (left). Due to low sequencing depth, the Frank et al. survey could have detected only 13 of the features considered discriminatory for classification (right vertical line).

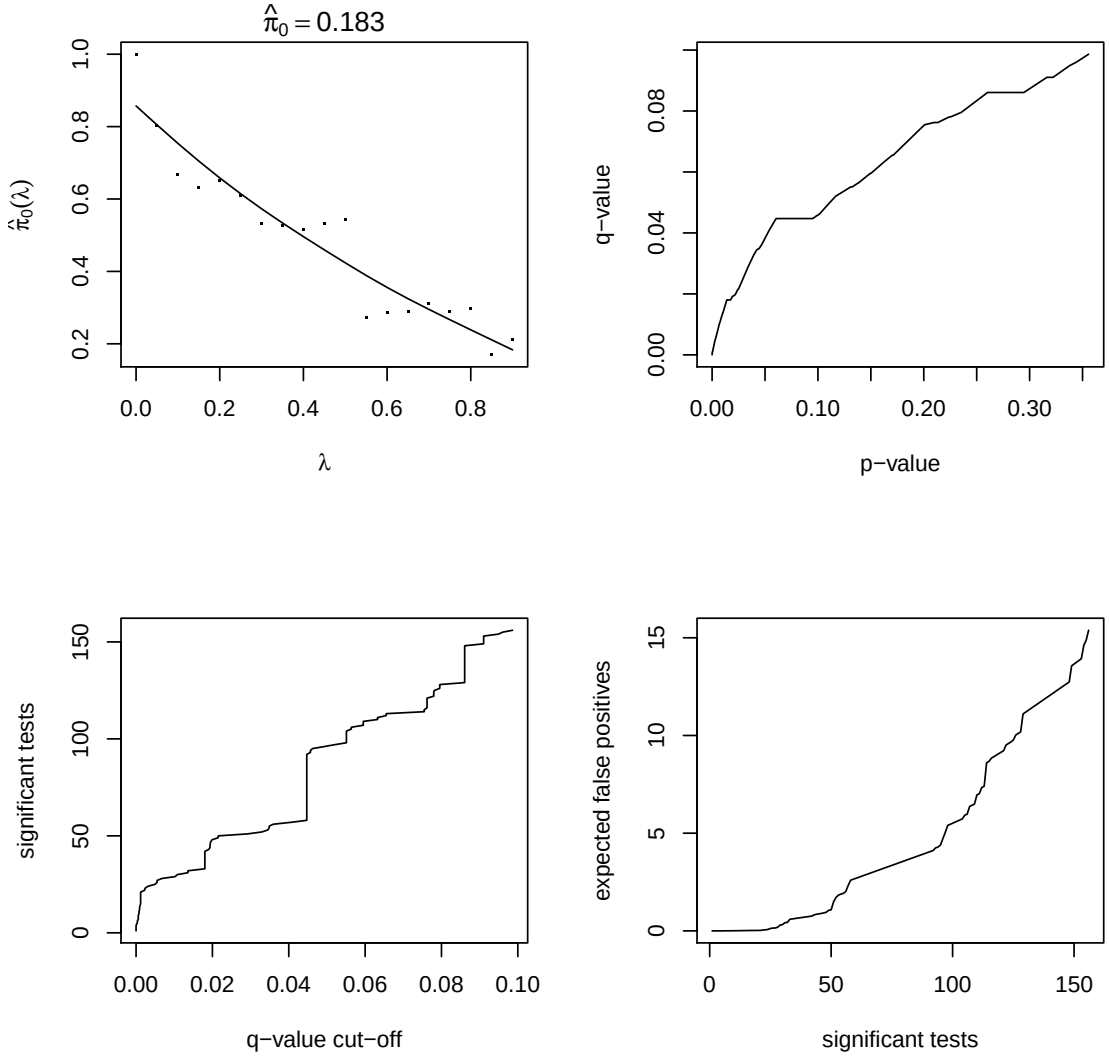

Figure 2.6: FDR adjustment of Kruskal-Wallis p-values for those features which best discriminate between active IBD samples and control samples. (Top-left) The expected proportion of false positive samples (po) is estimated by fitting. (Top-right) A plot of the calculated q-values versus the initial p-values. (Bottom-left) The number of significant tests for every given q-value cut-off. (Bottom-right) The number of expected false positives for a given number of significant tests considered.

genera decrease in the IBD patients with clinically active disease, but are still present in some IBD patients with inactive disease). This highlights the need for quantitative global surveys of microbial diversity rather than simple indicators of presence/absence.

Microbial alterations are similar in stool and tissue samples

Our finding that classification was similarly accurate in tissue and stool samples led us to ask whether the same alterations in the gut profile were observed in both sample types or whether distinct but similarly predictive changes occurred in each. To test this, we used the bacterial taxa identified in the pediatric case-control (stool-based) study to re-classify the tissue samples in the study by Frank et al. [7] The classification accuracy based on features from the pediatric study was nearly identical to the model using features picked from the tissue-based study: AUC = 0.71 (Figure 2.2), an increase in estimated measurement error of only 3%.

The relative change (upwards or downwards) of taxa in IBD vs. control groups is remarkably concordant between the two studies, with the exception of Lactobacillales (Figure 2.8). Unsurprisingly, due to the largely different sequencing depth many of the low-abundance taxa detected in the pediatric case-control (e.g., Alistipes) are of little importance in the classification, when applied to the Frank et al. data (Figure 2.9).

On the other hand, the Subdoligranulum genus and the Proteobacteria phylum remain two consistently important features across data sets (Figure 2.8). These results are encouraging because they suggest stool samples can be used to study changes in other compartments such as the mucosa.

Microbiota diversity decreases as disease severity increases

An important clinical question is to establish whether a marker of disease activity exists, and to what extent it can be used to stratify patients according to disease severity. To address this question, we measured disease activity by means of standard clinical indices (PUCAI, [88] PCDAI [89]), based on symptoms and blood test results, [90] and compared to SLiME predictions. While SLiME

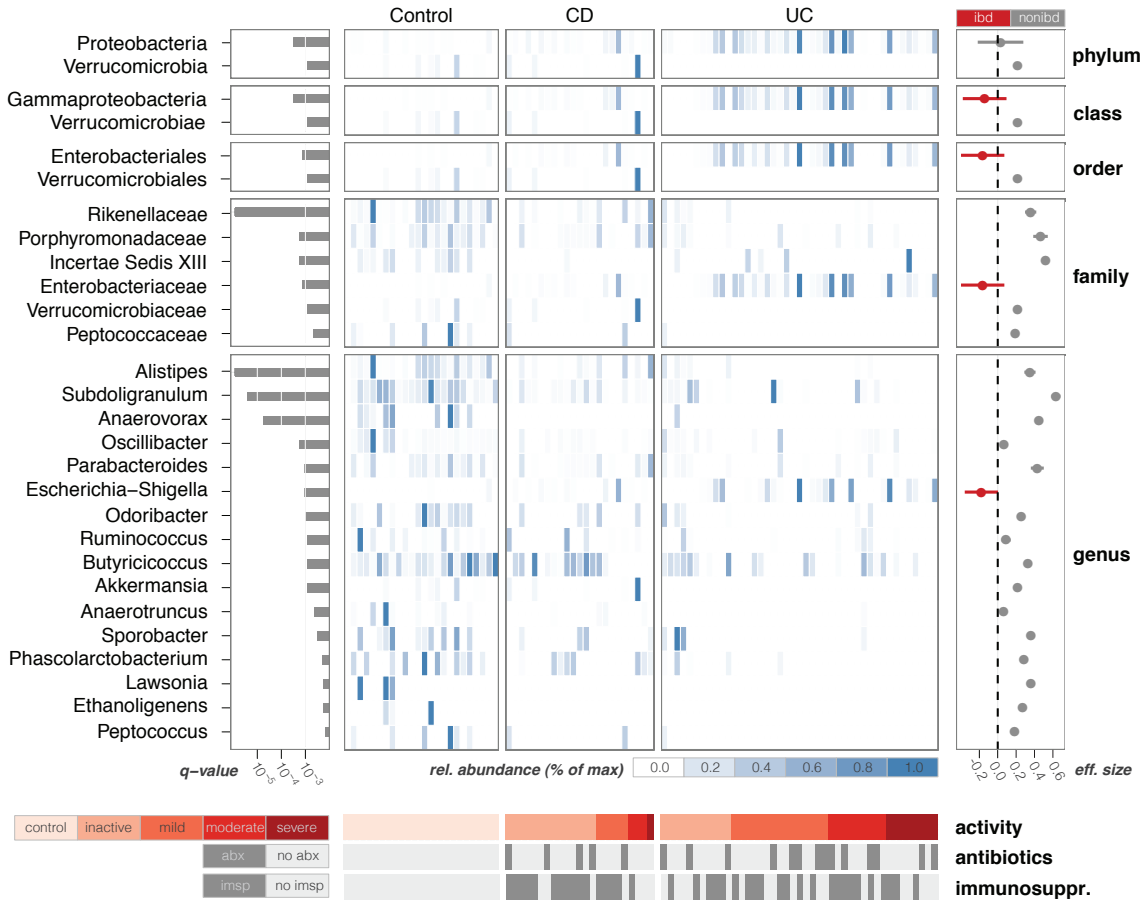

Figure 2.7: Taxa significantly associated with IBD. Center panel is a compositional heatmap of the selected taxa for each of the samples in the pediatric case-control study. Left panel indicates the significance of the association of each taxa with disease state, as measured by the q-value. Right panel shows a measure of effect size (Cohen's delta), highlighting in red those taxa which are significantly more prevalent in IBD samples. Bottom panels show relevant metadata for each sample, including disease activity as measured by PUCAI [88] and PCDAI indices [89].

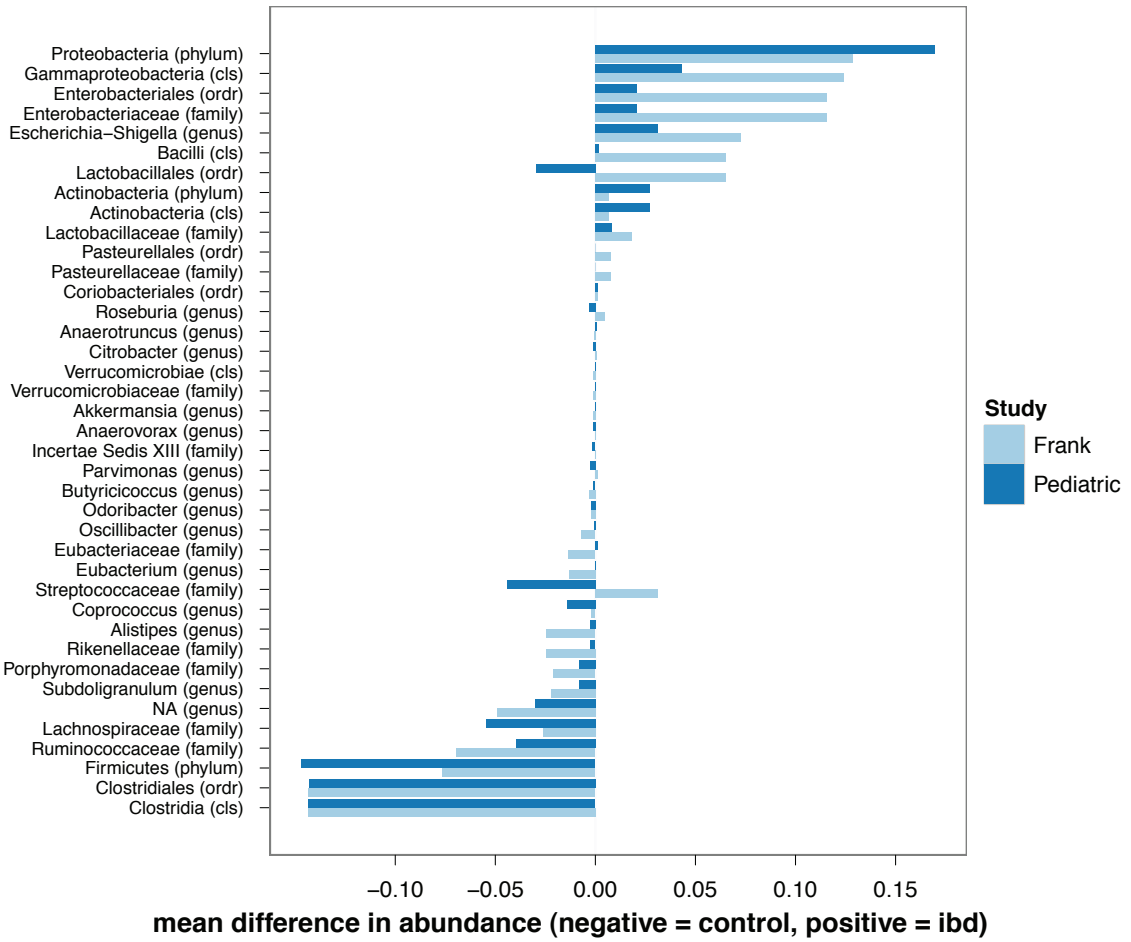

Figure 2.8: Taxa in the pediatric data set (stool-based) and the Frank et al. data set (tissue-based) agree in their relative abundance. Mean difference in normalized abundance between IBD samples and control samples is plotted for each taxa. Positive values (x-axis) mean the taxa is more prevalent in IBD samples, while negative values are associated with taxa more abundant in control samples. Stool-based and tissue-based data set are differentially colored (dark blue and light blue respectively).

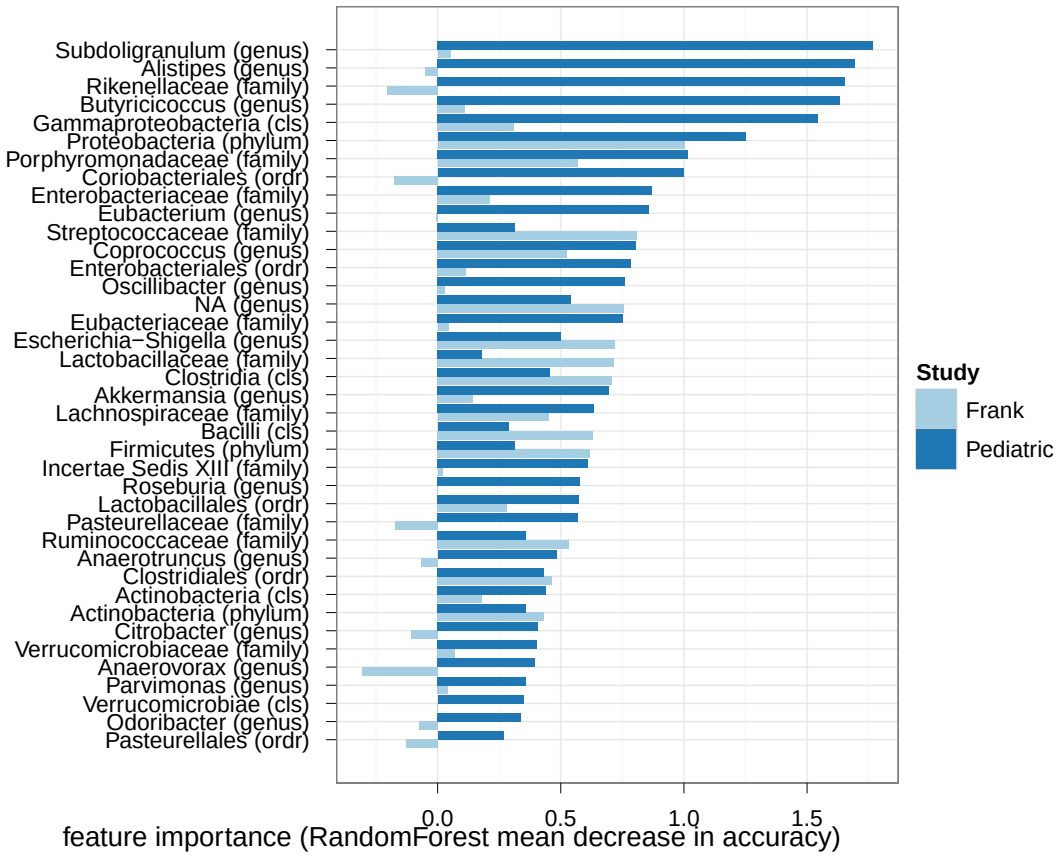

Figure 2.9: Taxa in the pediatric data set (stool-based) and the Frank et al. data set (tissue-based) vary in their importance as features. Best features - as determined by the RandomForest algorithm - applied to the pediatric data set are used to classify the Frank et al. data set. The importance of each feature - calculated as the decrease in accuracy of the algorithm when the feature is not used - is plotted for both studies. Noticeably, feature at the genus level are far more important in the pediatric data set than when used on the Frank et al. data set. This may reflect the greater depth of sequencing (see supplementary figure 2).

could not reliably classify on the basis of activity due to the small number of patients in each distinct level of disease severity, we nevertheless observed that overall microbiota diversity was strongly associated with disease activity. As disease severity increased, independently of the type of disease (CD or UC), overall bacterial diversity decreased as measured by the Shannon diversity index (Figure 2.10).

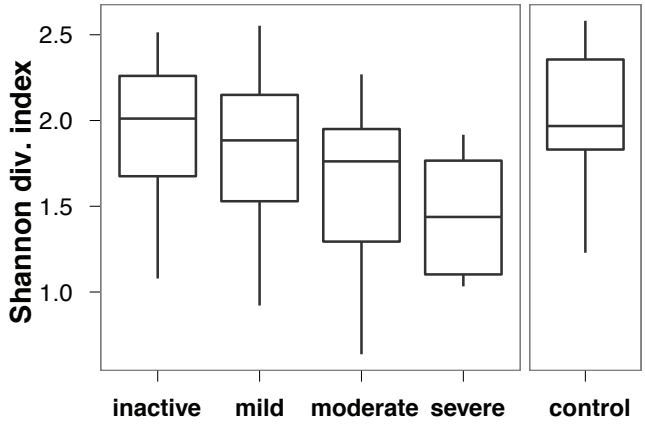

Figure 2.10: Stratification of patients by activity levels. (A) Overall microbial diversity as measured by the Shannon Diversity Index. Activity was assessed on the basis of patient symptoms using PCDAI and PUCAI clinical indices.

These results further support the view that IBD reflects an overall GI tract dysbiosis rather than the effect of a small number of pathogenic taxa. [77] [7] [91] Moreover, a number of microbial taxa showed significant association with disease activity levels. Among the most discriminative taxa was the Proteobacteria phylum (Figure 2.12 and Figure 2.11).

Specifically, the Gammaproteobacteria class was prevalent in all active forms of the disease. Severe disease in particular was associated with the Serratia and Escherichia-Shigella genera as well as the Coryneobacteriacea family.

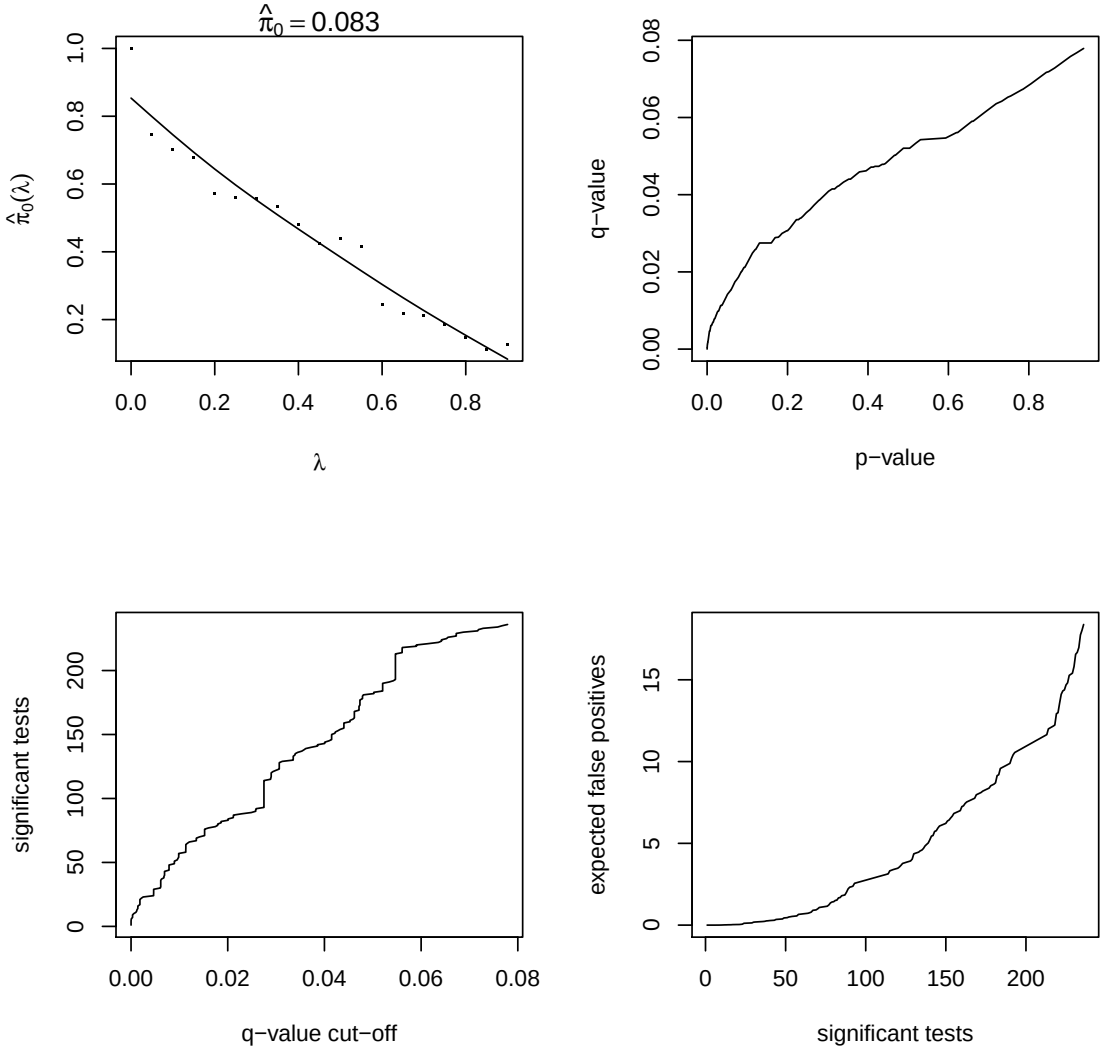

Figure 2.11: FDR adjustment of Kruskal-Wallis p-values for those features which best discriminate between levels of IBD activity. (Top-left) The expected proportion of false positive samples (po) is estimated by curve fitting. (Top-right) A plot of the calculated q-values versus the initial p-values. (Bottom-left) The number of significant tests for every given q-value cut-off. (Bottom-right) The number of expected false positives for a given number of significant tests considered.

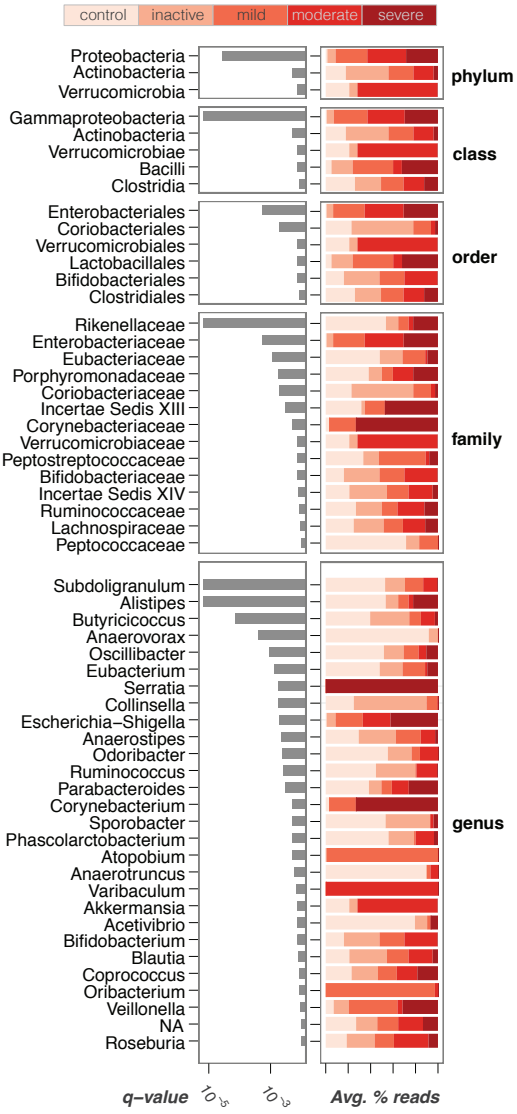

Figure 2.12: Best features to discriminate by activity levels. Activity levels are considered simultaneously, employing the Kruskal-Wallis test. Grey bar indicate the q-value and thus the strength of the association between the features and the disease state. Color bars indicate the average percentage of reads for each disease activity level.

Gut microbiota shows characteristic changes from active disease to remission

The factors responsible for triggering episodes of active disease are largely unknown. To identify microbial groups potentially associated with the establishment of active disease, we compared the composition of bacteria in fecal samples taken during active disease and remission periods. Classification with SLiME yielded significant differences between active and remission, with an AUC of 0.72. Amongst the taxa which were significantly associated with active disease (Figure 2.13) we found Proteobacteria (q-value $<$ 0.05, Kruskal-Wallis test, FDR adjusted, [87] E(π0)=0.35, see Figure 2.14) which was in agreement with previous observations. [92]

This finding appears to confirm the hypothesis that before or during active disease Proteobacteria rapidly expand and potentially displace other bacterial groups, such as Actinobacteria. On the other hand, members of the Eubacteriaceae, Incertae Sedis XIV and Bifidobacteriaceae families were associated with remission, which to our knowledge has not been reported previously. The Lachnospiraceae family, Subdoligranulum and Butyricicoccus, a butyrate-producing organism that can ferment dietary polysaccharides, were also associated with remission.

We hypothesized that individual-to-individual variation might obscure some of the signals of disease activity. To test this, we compared five patients at different stages of their disease. Consistent with our previous observations, we found an increase in Proteobacteria from samples taken during active phases of the disease in all five patients (see Figure 2.15).

In addition we found that Firmicutes and Actinobacteria phyla were distinctively prevalent during periods of remission for these patients. Also associated with remission were the genera Leuconostoc (heterofermentative bacteria capable of synthesizing dextran from sucrose), Roseburia (saccharolytic, butyrate-producing bacteria) and Eggerthella (anaerobic gram-positive bacilli).These results suggest that some disease-taxon associations may only be apparent when sampled are compared against an appropriate individual-specific baseline, highlighting the need for longitudinal studies of disease progression.

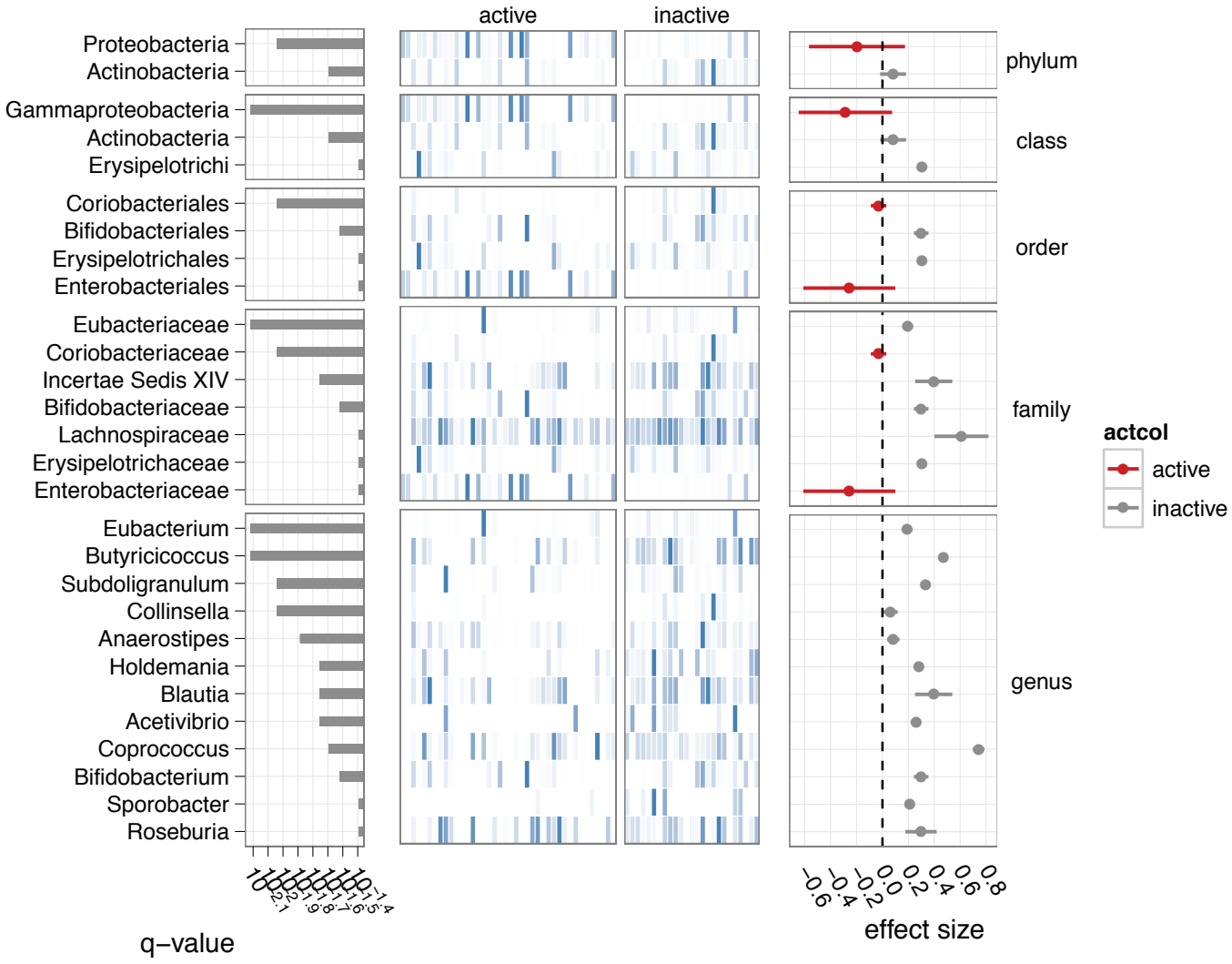

Figure 2.13: Features that show the greatest difference between active and inactive state in the pediatric case-control study. All features with significant association (q-value < 0.05, see supplementary figure 12) [87] to either active disease or remission are shown. Grey bars indicate the q-value of each taxon, heat maps describe the median normalized abundance in each sample. The right panel indicates the effect size and highlights in red the taxa which are prevalent in active samples.

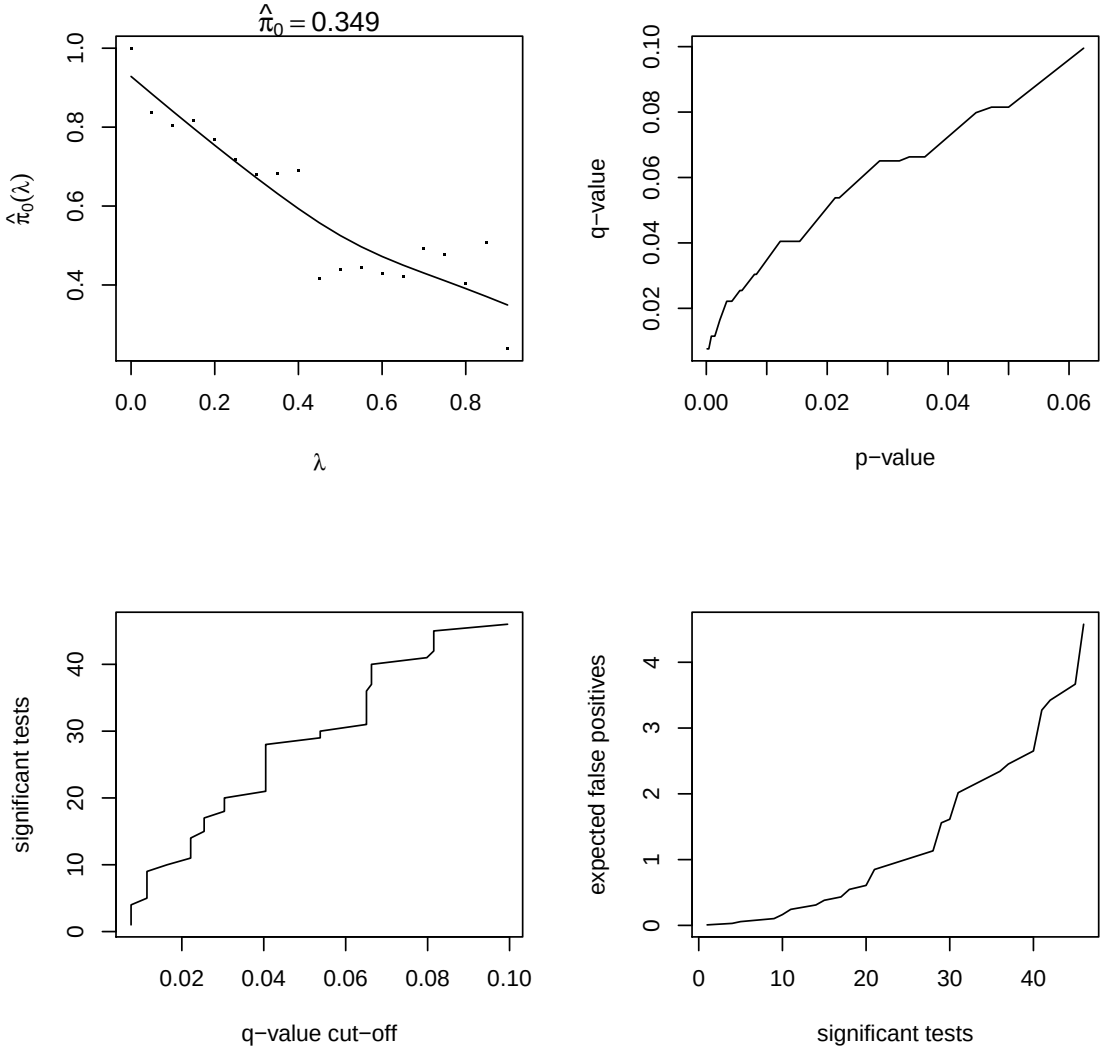

Figure 2.14: FDR adjustment of Kruskal-Wallis p-values for those features which best discriminate between active IBD samples and inactive IBD samples. (Top-left) The expected proportion of false positive samples (po) is estimated by curve fitting. (Top-right) A plot of the calculated q-values versus the initial p-values. (Bottom-left) The number of significant tests for every given q-value cut-off. (Bottom-right) The number of expected false positives for a given number of significant tests considered.

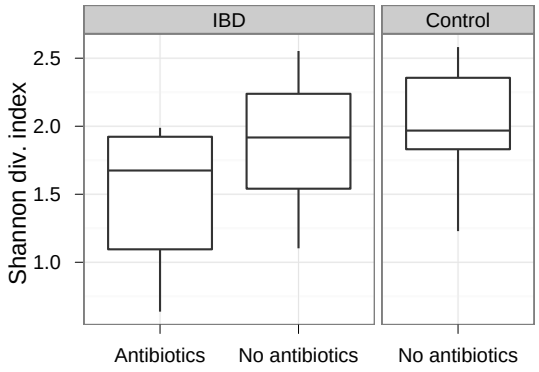

Figure 2.15: Antibiotic therapy reduces overall microbial diversity. Box plot showing the distribution of Shannon diversity indices for all patients undergoing antibiotic therapy, compared to the patients with IBD and without antibiotics, as well as controls.

Diversity is correlated with antibiotic therapy

We found that overall microbial diversity, as measured by the Shannon diversity index, was the single most important feature for discriminating between patients undergoing antibiotic therapy or not. Although we could not classify whether samples were obtained from antibiotic-treated patients with high accuracy (AUC $<$ 0.6), we did find that Shannon diversity index was significantly and negatively associated with antibiotic therapy in the IBD samples (p-value = 0.0067, Wilcoxon test, see Figure 2.16).

This observation is consistent with a simple model of antibiotic effect on the gut microbiota: most taxa and bacterial groups are killed by antibiotics, while the few bacterial strains which have resistance survive and increase in relative abundance.

Differential diagnosis of ulcerative colitis and crohn's disease is possible

Ulcerative colitis is generally limited to the colon, while intestinal inflammation in Crohn's disease may occur in any region of the gastrointestinal tract. Classification of pediatric IBD patients into UC or CD at the time of fecal testing is desirable, given the different clinical management of the two

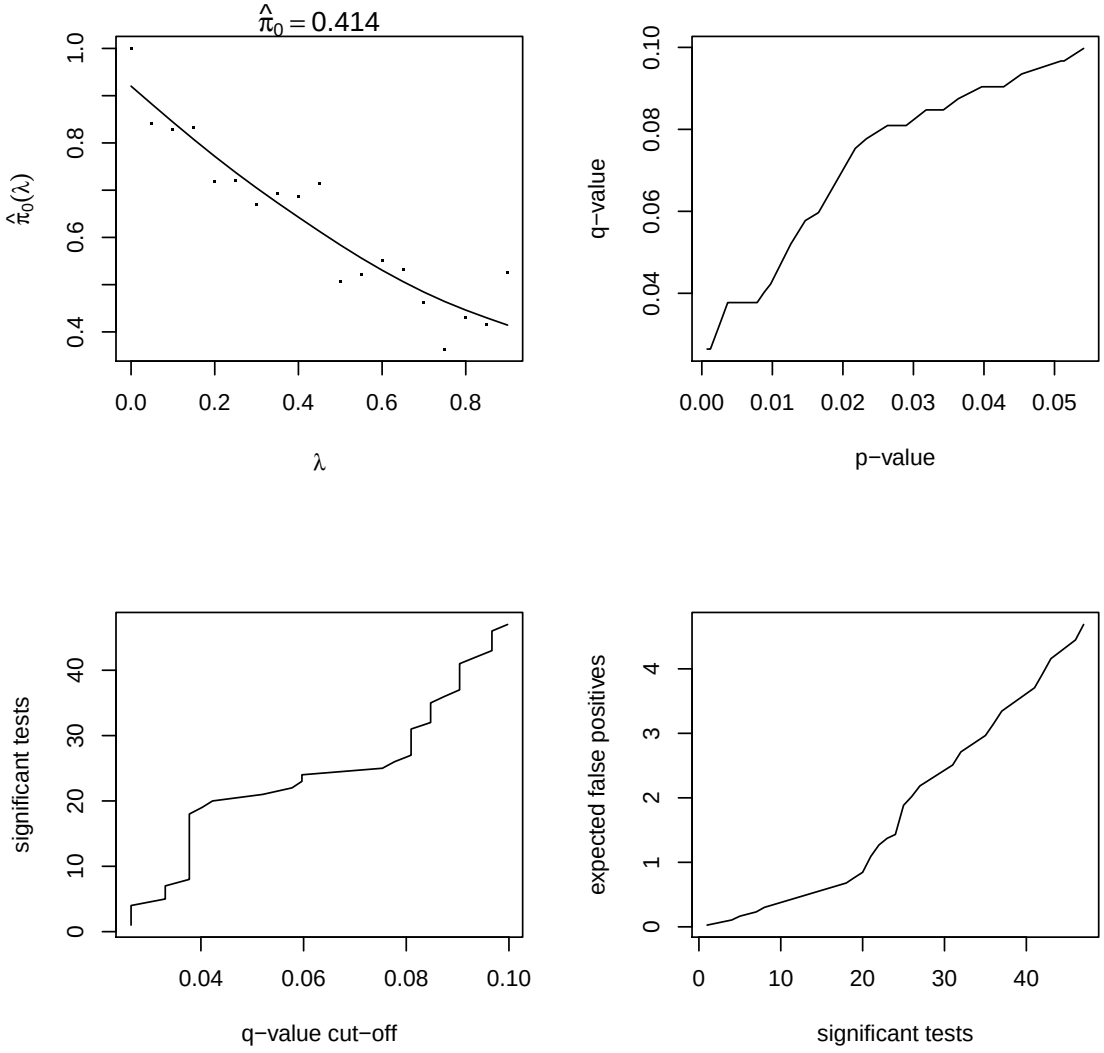

Figure 2.16: FDR adjustment of Kruskal-Wallis p-values for those features which best discriminate between CD samples and UC samples. (Top-left) The expected proportion of false positive samples (po) is estimated by curve fitting. (Top-right) A plot of the calculated q-values versus the initial p-values. (Bottom-left) The number of significant tests for every given q-value cut-off. (Bottom-right) The number of expected false positives for a given number of significant tests considered.

diseases. Even though distinguishing UC from CD was not the primary aim of our study design, we found that SLiME applied to the case-control data set could separate UC patients and control patients accurately (Figure 2.17, cross-validated AUC=0.82 and 0.83 respectively), but was less accurate in distinguishing Crohn's disease patients (AUC=0.68). When we excluded controls from the data and attempted to distinguish between CD and UC in all IBD patients, we were able to do so with accuracies (AUC=0.76,ie. a specificity of 49% at 95% sensitivity) superior to current noninvasive clinical methods [72].

The most informative bacterial families in discriminating UC, CD and Control samples as determined by Kruskal-Wallis test were the Eubacteriaceae, Bacteroidaceae, and Verrucomicrobiaceae (Figure 2.19 and Figure 2.18).

Verrucomicrobia were consistently employed in the classifier because bacteria of this group were completely absent from UC patients, which tended to be characterized by Lactobacillales or Bacilli and Gammaproteobacteria.

Steroid treatment could potentially affect the composition of the microbiota and in turn the accuracy of the classification between CD and UC. To assess this effect, we limited our analysis to those patients undergoing steroid therapy. However, we found no substantial difference in the accuracy of the classification (AUC = 0.73, 40% specificity at 95% sensitivity) between CD and UC patients in the steroid subgroup with respect to the totality of all IBD patients.

Classification in CD or UC performed differently depending on whether the patient was experiencing active disease or remission, and surprisingly was more accurate at distinguishing CD and UC patients in remission (AUC=0.73) than for patients with active disease (AUC=0.67). This finding suggests that changes in microbiota composition during acute inflammation may be similar in both UC and CD, rendering distinction by microbial diversity more challenging.

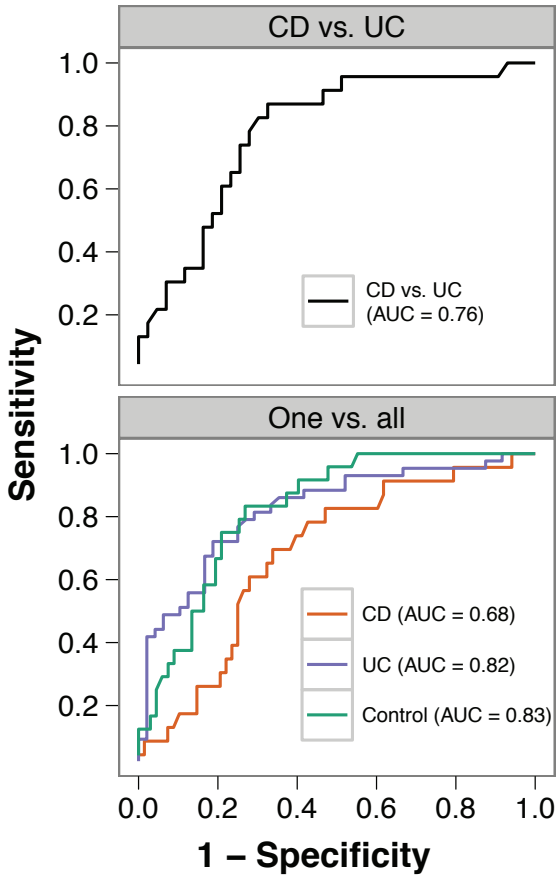

Figure 2.17: Discrimination of CD and UC. Above, ROC curve for the classification of CD vs UC in samples where diagnosis of IBD is already established. Below, ROC curve for the classification of each disease class against all other classes.

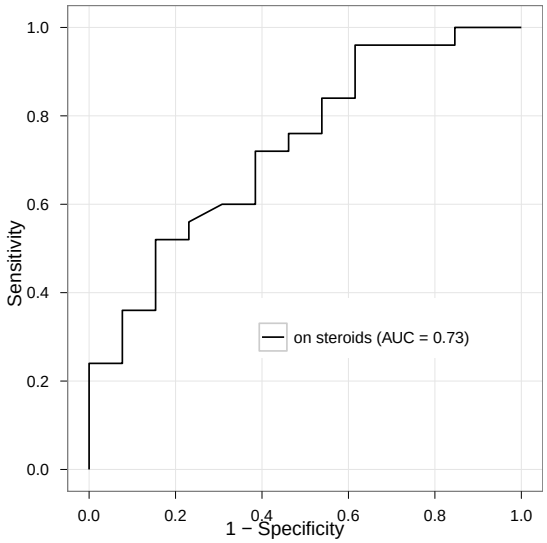

Figure 2.18: ROC curve for the CD vs UC classification in the steroid-treated subgroup. The performance in this subset of the cohort is comparable to the totality of IBD patients.

Blind validation with an independent patient sample confirms the accuracy of supervised classification

To confirm the general validity of our results, we selected an independent patient sample of 77 children and young adults. Following fecal sampling and 16S rRNA sequencing, we applied SLiME – trained on our initial pediatric cohort – to the new dataset. Encouragingly, SLiME maintains good performance in distinguishing IBD patients from controls (AUC = 0.74, Figure 2.20).

Table 2.3 illustrates the classification performance of SLiME on the validation cohort when the model is tuned for high sensitivity (91.6%) and low specificity (17.6%).

Classification by SLiME is comparable to testing by fecal calprotectin

We compared the accuracy of SLiME with the outcome of the fecal calprotectin test on a portion of our samples from both the pediatric cohort and the validation cohort, to determine how our method compared to the most clinically accepted non-invasive test for IBD. A summary of the re-

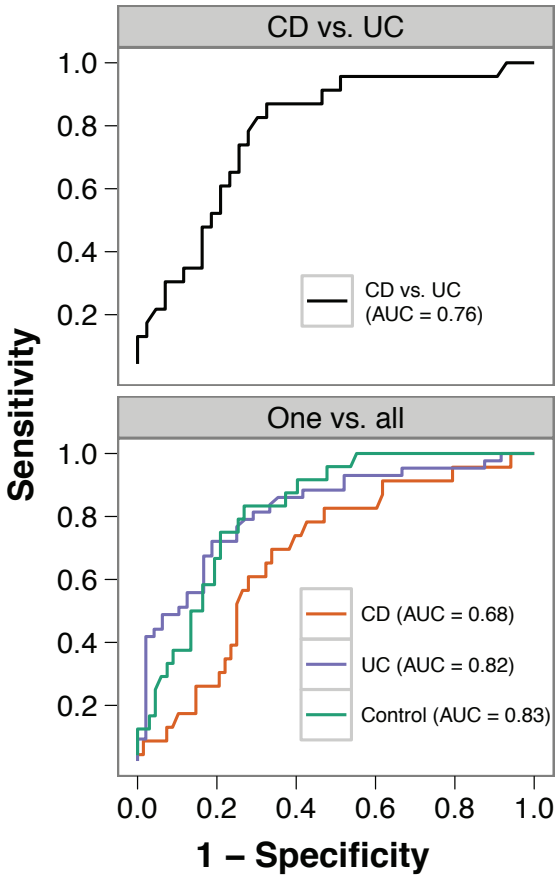

Figure 2.19: Strength of association for the best features (q-value < 0.05) [87] which allow discrimination between CD and UC.

Table 2.3: Confusion matrix for the blind validation of the SLiME classifier on an independent validation cohort. Sensitivity for IBD vs controls is 94.5% while specificity is 46.1%. Note this is only one possible cutoff value. Different sensitivity and specificity can be obtained by appropriately tuning the cutoff.

	SLiME classification	
Diagnosis	IBD	non-IBD
CD	24	2
UC	28	1
Control	7	6

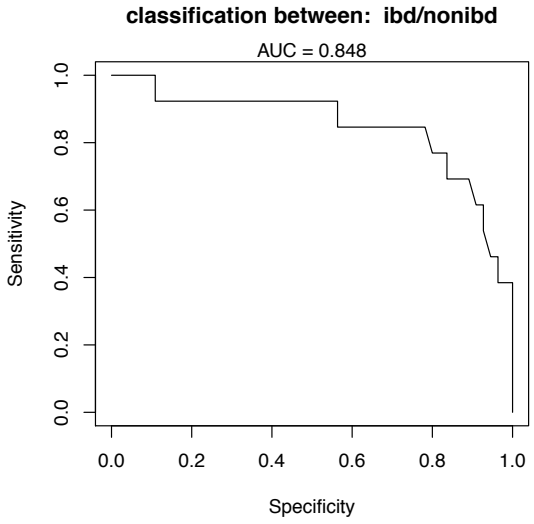

Figure 2.20: Blind validation of a SLiME model — previously trained on our pediatric cohort — applied to an independent set of fecal samples from 77 patients. ROC curve shows that high sensitivity and high specificity are maintained across a range of disease prevalences.

sults is contained in Table 2.4 and Table 2.5. On those samples where we could obtain calprotectin measurements retrospectively, we found that SLiME could classify the samples as IBD with comparable accuracy to calprotectin (AUC = 0.79 compared to calprotectin's AUC of 0.81). Superposing the two ROC curves (Figure 2.21 and Figure 2.22) shows that calprotectin was only slighly more sensitive than SLiME.

Table 2.4: Summary of calprotectin assay results per disease condition; Assay range: 10 mcg/g - 600 mcg/g

Diagnosis	n	Mean	St.dev
CD	36	444.6	241.2
UC	54	467.6	223.6
Control	29	196.4	208.2

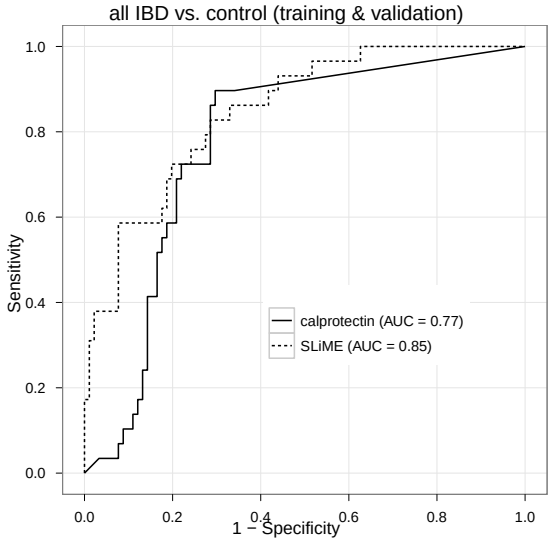

Figure 2.21: Comparison of SLiME and fecal calprotectin assay. The two assays have comparable efficacy in distinguishing IBD patients from control when applied to all samples in the training and validation cohorts for which calprotectin could be measured (n=120)

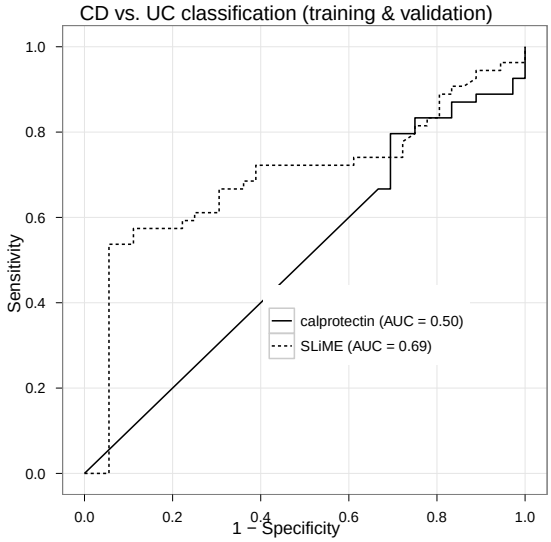

Figure 2.22: Comparison of SLiME and fecal calprotectin assay. SLiME is slightly superior in distinguishing CD from UC samples, when applied to all CD and UC samples (n=90) in the training and validation cohorts for which calprotectin could be measured

Table 2.5: Summary of calprotectin assay results per activity level; Assay range: 10 mcg/g - 600 mcg/g

Activity	n	Mean	St.dev
Active	48	501.1	205.6
Inactive	43	401.3	251.29
Control	29	196.4	208.2

2.3 Discussion

Delay in the diagnosis of pediatric IBD is likely due to the non-specific presentation of the disease. An inexpensive and sensitive diagnostic tool could reduce this delay by rapidly identifying patients at high risk for IBD and, therefore, warranting endoscopic evaluation. In this study, we demonstrated the feasibility of a new approach to detecting pediatric IBD based on analysis of fecal microbiota. The sensitivity and specificity of our approach, as measured by ROC curve analysis, matches or surpasses that of alternative methods proposed for clinical use.

Two key methodological advances are responsible for improved performance compared to previous studies. These include the SLiME software package, which is freely available for public use, and increased sampling depth, which allows low abundance but highly informative groups to be sampled. The advantages of employing machine learning methods to analyze microbiome data have already been discussed. [84] Compared to clustering methods, machine learning excels in classifying unlabelled data and extracting pivotal features from large and complex microbiome data sets.

Previous surveys of microbial diversity in IBD relied on clustering analyses to differentiate between IBD and non-IBD samples. [7] [92] [76] As a result, these studies suffered from poor sensitivity and, more importantly, did not generate predictive models that could be employed to distinguish new unlabelled samples. In this study, we employed SLiME to achieve high sensitivity as well as high specificity in differentiating active IBD samples from controls. Models generated by SLiME were capable of classifying unlabelled samples with accuracy, as demonstrated by the large AUC obtained both after cross-validation and after blind validation with an independent cohort.

Importantly, our approach was effective across disparate data sets using different sample types, and processing and sequencing technologies. Finally, SLiME generates a list of taxa specifically associated with each disease state (active IBD, remission samples, CD and UC) facilitating biological interpretation.

Although we succesfully employed specific taxa as predictive biomarkers, our results indicate that IBD reflects an overall GI tract dysbiosis rather than the effect of a small number of pathogenic taxa. This result is in agreement with previous observations [77] [7] [91] and suggests that a global community survey rather than a test for bacterial presence/absence is better suited to identifying IBD.

Departing from the traditional clustering analysis, a recent and promising study [81] showed the use of a predictive model in classifying samples as IBD on the basis of microbial diversity. However, the same study arised concerns regarding a) the ability to distinguish UC patients from controls and b) the ability to discriminate between samples from patients with active disease and those in remission. Our study answers these questions, and importantly we report only cross-validated results that should more closely reflect accuracy in a clinical setting.

Some potential limitations in our study stem from its relatively small scale. For instance, while we are able to succesfully distinguish both UC and CD patients, SLiME appears to classify UC patients more succesfully than CD patients. However, we find that this difference in performance disappears after downsampling, confirming that it is due to the uneven split between CD and UC patients in our training cohort. We also attempted to find correlations between therapeutic regimens (antibiotics, salicylates, anti-TNF, methotrexate, etc.) and microbial composition. Unfortunately SLiME was not capable to differentiate between subgroups with different therapeutic regimens, most likely due to the broad range of treatments employed in our cohort and the small number of patients in each subgroup. While these results indicate that SLiME may not be influenced by different therapeutic interventions while differentiating patients with IBD from controls, recruiting a larger number of patients following similar therapeutic regimens would have allowed to identify key microbial changes brought about by the therapy.

It is arguable that both these potential limitations will be addressed by studies with larger patient samples, better suited to compare alternatives in disease behaviour and therapeutic management of IBD. In addition, a cross-sectional study design on fecal samples taken at the time of diagnosis and before the start of any therapy, rather than the case-control study we employed, would allow to estimate more precisely the sensitivity of SLiME when employed in the general population.

Even though our results demonstrate the potential of the gastrointestinal microbiome as a diagnostic tool in IBD, further validation will be necessary before this tool is accepted into clinical practice. Our comparison between SLiME and calprotectin is encouraging, insofar as it shows that the two methods have comparable accuracies on this data set. Further comparison of SLiME against other IBD biomarkers, including C-reactive protein, fecal lactoferrin, and fecal calprotectin [69] in larger patient samples will allow clinicians to gauge the relative benefits of each method.

Despite these limitations, our results demonstrate the considerable potential of microbiome-based diagnostics in the clinic, particularly in the case of pediatric patients where diagnosis is often challenging. Similar approaches could evaluate the efficacy of novel therapies (e.g. probiotics, antibodies), predict the outcome of disease and forecast the timings of flare-ups. While not replacing endoscopy and histological examination as diagnostic tools, we propose that classification based on microbial diversity can be included as an effective complementary technique to aid in the diagnosis of IBD, particularly in pediatric patients.

The aim of efforts to improve diagnostics for IBD revolve around the central tenet of avoiding invasive techniques such as endoscopy, particularly in pediatric patients for whom these procedures can be particularly disagreeable. SLiME, in its current form, is not sensitive enough to unequivocally distinguish IBD; neither are currently used serological tests. However, with certain additional improvements to these methodologies, they can potentially be combined to significantly improve sensitivity.

According to the internationally mandated guidelines on IBD, suspected cases should be subjected to a full blood count, an erythrocyte sedimentation rate, liver function tests and a measure

of C reactive protein. In addition, a test of calprotectin is often included which can quantify inflammation and is somewhat specific to the gut. These values do not usually result in an unambiguous diagnosis, and even when the indications of IBD are strong, they are still unable to differentiate between CD and UC. Perinuclear antineutrophil cytoplasmic antibody and anti-saccharomyces cerevisiae antibody are positively associated with UC and CD respectively, but with low sensitivity and with little clinical utility. In addition, many other conditions can confound diagnosis of IBD and must be excluded in suspected cases. These include diarrheal diseases and certain gut parasitic infections. With the plummeting costs of sequencing, one can imagine including microbial profiling to the panel of stool tests. SLiME classification will then be able to yield a likelihood of the stool sample being result of IBD and thus suggest earlier the need for a colonoscopy.

The main strengths of SLiME as applied to IBD patient cohorts are its non-invasiveness and its ability to not only identify IBD but to also distinguish between CD and UC. Further training of the algorithm on additional datasets, as they become available, stands to improve the sensitivity of its predictive power and may ultimately replace the need for colonoscopy, at least in children. Although with the current training data it cannot be used as a stand-alone test, it can easily be combined with the currently accepted serology, particularly as cost of microbial shotgun sequencing will decrease. Even if each test on its own is not categorical, the diagnostic capability of them combined stands to be much superior. Moving forward, it will be important to test SLiME as a component of a multi-pronged non-invasive diagnostic approach for this disease.

2.4 Methods

Patient population and recruitment

Fecal samples were obtained from 91 children and young adults with Crohn's disease, ulcerative colitis, and non-inflammatory conditions of the gastrointestinal tract. The control population was composed of patients with symptomatology suggestive of IBD: constipation (n=9), abdominal

pain (n=8), gastroesophageal reflux (n=2), poor weight gain (n=1), diarrhea (n=1), blood in stool (n=2) and oropharyngeal dysphagia (n=1) (see Table 2.6). Table 2.7 shows the patient demographics. Recruitment was conducted in the clinic or inpatient hospital wards under a protocol approved by the Children's Hospital Committee on Clinical Investigation.

Table 2.6: Control patients' diagnoses

	Training Cohort (n = 24)	Validation Cohort (n = 13)
Functional Abdominal Pain	8 (33%)	4 (30%)
Constipation	10 (42%)	1 (8%)
Irritable Bowel Syndrome	0	3 (23%)
Vomiting	0	3 (23%)
Rectal Bleeding with Normal EGD/Colon	1 (4%)	1 (8%)
Diarrhea	2 (8%)	0
Reflux	1 (4%)	0
Poor Growth	1 (4%)	0
Feeding Difficulty	1 (4%)	0
History of Food Allergies	0	1 (8%)

Fecal samples were generally obtained within 4 hours of the bowel movement, and stool was frozen at –70 degrees C on the receipt of the sample from the patient. Clinical data and possible confounding variables were recorded at the time of sample acquisition including: disease type, disease location, disease duration, disease activity (as determined by the Pediatric Crohn's disease activity index for CD, and the pediatric ulcerative colitis activity index for UC) [89] [88], and current prescribed medications. An additional independent patient sample of 68 children and young adults was selected for blind validation. Table 2.8 shows the patient demographics of this additional sample set.

Histological evidence and data on disease duration for both sample sets are contained in Table 2.11, Table 2.12, Table 2.9 and **??**.

Table 2.7: demographics of paediatric (training) cohort

		Crohn (n=23)	UC (n=43)	Control (n=24)	IBDU (n=1)
Gender					
	Male	13(56%)	21(49%)	10(42.%)	1
	Female	10(44%)	22(51%)	14(58%)	0
	Age (Median +/-)	14.13 ± 3.84	13.7 ± 4.25	9.08 ± 4.3	14
		Range	3–20	4–24	3–17
	Montreal Classification				
	L1	1(4%)			
	L2	1(4%)			
	L3	15(65%)			
	L4	6(26%)			
	B1	18(78%)			
	B2	4(17%)			
	B3	1(4%)			
	E1		2(5%)		
	E2		6(14%)		
	E3		35(81%)		
Disease Activity					
	Control	0	0	24	0
	Inactive	14	11		
	Mild	5	15		1
	Moderate	3	9		
	Severe	1	8		
	Medications				
	Salicylates only	1 (4%)	6(14%)		
	6mp/AZA/MTX	11 (48%)	12(28%)		
	Anti-TNF	7 (30%)	4(9%)		
	Calcineurin inhibitor	0	23(53%)		
	Antibiotics	6 (26%)	14(33%)		1
	(Steroids)	13 (57%)	25(58%)		1

Table 2.8: Patient demographics for the validation set

		Crohn's (n=25)	Control (n=13)	UC (n=30)
	Gender			
	Male	16 (64%)	6 (46%)	18 (60%)
	Female	9 (36%)	7 (54%)	11 (60%)
Age				
	Median +/-	15 +/- 4.7	13 +/- 6.8	13 +/- 4.0
	Range	5–23	3–21	5–21
Montreal Classification				
	L1	3 (12%)		
	L2	0		
	L3	7 (28%)		
	L4	1 (4%)		
	B1	16 (64%)		
	B2	1 (4%)		
	B1p	6 (24%)		
	B2p	1 (4%)		
	B3p	1 (4%)		
	L1 + L4	1 (4%)		
	L2 + L4	4 (16%)		
	L3 + L4	9 (36%)		
	E1			23 77%)
	E2			6 (20%)
	E3			1 (3%)
Disease Activity				
	Control	0	13	0
	Inactive	15 (60%)		15 (50%)
	Mild	6 (24%)		7 (23%)
	Moderate	2 (8%)		6 (20%)
	Severe	2 (8%)		2 (7%)
Medications				
	Salicylates only	1 (4%)	0	11 (37%)
	6mp/AZA/MTX	11 (44%)	0	12 (40%)
	Anti-TNF	5 (20%)	0	2 (7%)
	Calcineurin inhibitor	0	0	6 (20%)
	Antibiotics	4 (16%)	0	13 (43%)
	Steroids	9 (36%)	0	11 (40%)

Table 2.9: Disease Duration at Time of Sample Acquisition. Initial (Training) Cohort (n = 91)

	Median Disease Duration (months)	Mean Disease Duration (months)	Interquartile Range (months)
All IBD (n = 67)	21	34.8	58.0 (4.0 - 62.0)
Crohn's Disease (n = 23)	28	38.4	67.0 (2.0 - 69.0)
Active Disease (n = 9)	5	29.6	69.0 (0.3 - 69.3)
Steroids (n = 5)	0.5	2.25	6.0 (0.3 - 6.3)
Antibiotics (n = 0)	-	-	-
Steroids & Abx (n = 2)	50.5	50.5	48 (26.25 - 74.75)
Inactive Disease (n = 14)	31.5	44	70.8 (5.0 - 75.8)
Steroids (n = 4)	23.5	35.8	82.8 (0.5 - 83.3)
Antibiotics (n = 1)	112	112	0
Steroids & Abx (n = 2)	65	65.0	63.0 (33.5 - 96.5)
Ulcerative Colitis (n = 43)	21	33.6	55.0 (5.0 - 60.0)
Active Disease (n = 32)	19.5	35.6	59.5 (4.3 - 63.8)
Steroids (n = 12)	6	20.8	28.3 (0.5 - 28.8)
Antibiotics (n = 2)	69.5	69.5	4.5 (67.3 - 71.8)
Steroids & Abx (n = 11)	10.5	35.4	50.2 (0.8 - 51.0)
Inactive Disease (n = 11)	26	32.4	44.0 (9.0 - 53.0)
Steroids (n = 2)	5	5.0	4.0 (3.0 - 7.0)
Antibiotics (n = 1)	80	80.0	0
Steroids & Abx (n = 0)	-	-	-
IBDU (n = 1)	4	4.0	0

Table 2.10: Disease Duration at Time of Sample Acquisition. Validation Cohort (n = 68)

	Median Disease Duration (months)	Mean Disease Duration (months)	Interquartile Range (months)
All IBD (n = 55)	22	33.0	48.0 (4.0 - 52.0)
Crohn's Disease (n = 25)	24	41.0	69.0 (4.0 - 73.0)
Active Disease (n = 10)	63	59.1	71.8 (12.0 - 83.8)
Steroids (n = 3)	53	57.2	58.7 (26.8 - 85.5)
Antibiotics (n = 2)	78	78.0	5.0 (75.5 - 80.5)
Steroids & Abx (n = 2)	4	4	0
Inactive Disease (n = 15)	18	28.9	32.5 (2.0 - 34.5)
Steroids (n = 4)	0.8	0.9	0.9 (0.4 - 1.25)
Antibiotics (n = 0)	-	-	-
Steroids & Abx (n = 0)	-	-	-
Ulcerative Colitis (n = 30)	20	26.3	39.0 (5.0 - 44.0)
Active Disease (n = 15)	20	22.6	35.9 (0.6 - 36.5)
Steroids (n = 1)	0.5	0.5	0
Antibiotics (n = 3)	46	53.0	27.5 (37.5 - 65)
Steroids & Abx (n = 8)	5	14.1	21.1 (0.7 - 21.8)
Inactive Disease (n = 15)	20	30.0	37.5 (9.5 - 47.0)
Steroids (n = 2)	10.5	10.5	9.5 (5.75 - 15.25)
Antibiotics (n = 2)	55	55.0	12.0 (49.0 - 61.0)
Steroids & Abx (n = 0)	-	-	-

Table 2.11: Histological evidence of disease at diagnostic colonoscopy. Initial (Training) Cohort (n = 91)

	Crohn's (n=23)	UC (n=43)	IBDU (n=1)
Ileal Disease			
Active Ileitis	16 (70%)	3 (7%)	1 (100%)
Ileal Ulceration	7 (30%)	0	0
Ileal Granulomas	4 (17%)	0	0
Colonic Disease			
Chronic Active Colitis	19 (83%)	43 (100%)	1 (100%)
Chronic Inactive Colitis	0	2 (5%)	0
Continuous Colitis	5 (22%)	42 (98%)	1 (100%)
Patchy Colitis	14 (61%)	1 (2%)	0
Colonic Granulomas	11 (48%)	0	0

Table 2.12: Histological evidence of disease at diagnostic colonoscopy. Validation Cohort (n = 68)

	Crohn's (n= 25)	UC (n= 30)	IBDU (n=0)
Ileal Disease			
Active Ileitis	19 (76%)	2 (7%)	0
Ileal Ulceration	9 (36%)	0	0
Ileal Granulomas	7 (28%)	0	0
Colonic Disease			
Chronic Active Colitis	17 (68%)	30 (100%)	0
Chronic Inactive Colitis	2 (8%)	6 (20%)	0
Continuous Colitis	8 (32%)	26 (87%)	0
Patchy Colitis	10 (40%)	4 (13%)	0
Colonic Granulomas	12 (48%)	0	0

DNA extraction and pyrosequencing

DNA from stool samples was extracted using the QIAamp DNA Stool Mini Kit (Qiagen, Inc., Valencia, CA) according to manufacturer's instructions. The manufacturer protocol was altered to accommodate larger volumes of stool and to improve homogenization using bead-beating techniques at several steps: a) a minimum of 2mL of Buffer ASL and 300mg of stool was used in the protocol; b) a ratio of 700uL of Buffer ASL per 100mg of stool weight was used for larger volumes using no more than 1500mg of stool and 10.5mL of Buffer ASL; c) following the addition of Buffer ASL to each sample (step #2), 0.70 mm Garnet Beads (MO BIO Laboratories, Inc., Carlsbad, CA) were added to the suspension and vortexed for 10 seconds; d) a second bead-beating was done following the heating of the suspension (step #3) in 0.1 mm Glass Bead Tubes (MO BIO Laboratories, Inc., Carlsbad, CA), and vortexed for 10 minutes.

Extracted DNA was employed for 454 FLX Titaninum pyrosequencing of PCR-amplified windows of the 16S gene. Variable region V3-V5 amplification primers were designed with FLX Titanium adaptors (A adaptor sequence: 5' CCATCTCATCCCTGCGTGTCTCCGACTCAG 3'; B adaptor sequence: 5' CCTATCCCCTGTGTGCCTTGGCAGTCTCAG 3') on the 5' end of the 16S primer sequence: 454B_357F (5' CCTACGGGAGGCAGCAG 3') and 454A_barcode_926R (5' CCGTCAATTCMTTTRAGT 3'). See Table 2.13. Polymerase chain reaction (PCR) mixtures (25µl) contained 10ng of template, 1x Easy A reaction buffer (Stratagene, La Jolla, CA), 200mM of each dNTP (Stratagene), 200nM of each primer, and 1.25U Easy A cloning enzyme (Stratagene). The cycling conditions for the V3-V5 consisted of an initial denaturation of 95°C for 2 min, followed by 25 cycles of denaturation at 95°C for 40 sec, annealing at 50°C for 30 sec, extension at 72°C for 5 min and a final extension at 72°C for 7 min. Amplicons were confirmed on 1.2% Flash Gels (Lonza, Rockland, ME) and purified with AMPure XP DNA purification beads (Beckman Coulter, Danvers, MA) according to the manufacturer and eluted in 25 µL of 1X low TE buffer (pH 8.0). Amplicons were quantified on Agilent Bioanalyzer 2100 DNA 1000 chips (Agilent Technologies, Santa Clara, CA) and pooled in equimolar concentration. Emulsion PCR and sequencing

were performed according to the manufacturer's specifications. Sequencing was performed with a target of 5000 raw reads per sample.

Sanger sequencing

Polymerase chain reaction (PCR) mixtures (25µl) contained 10ng of template, 1x Easy A reaction buffer (Stratagene, La Jolla, CA), 200mM of each dNTP (Stratagene), 200nM of each primer (63f: 5' GCCTAACACATGCAAGTC 3'; U1525R: 5' AAGGAGGTGWTCCARCC 3'), and 1.25U Easy A cloning enzyme (Stratagene). The cycling conditions consisted of an initial denaturation of 95°C for 2 min, followed by 30 cycles of denaturation at 95°C for 40 sec, annealing at 50°C for 30 sec, extension at 72°C for 2 min and a final extension at 72°C for 7 min. PCR products were purified with QIAquick PCR purification kit (QIAGEN, Inc, Valencia, CA) according to the manufacturer, and size selected on a 1% agarose gel. The gel bands were purified with QIAquick gel extraction kit (QIAGEN) according to the manufacturer's instructions with one modification: the gel bands were dissolved at room temperature on a Dynal Bioteck Rotator (Model RKDYNAL, setting 30, Invitrogen, Life Technologies, Carlsbad, CA) for 15 minutes. Cleaned amplicons were cloned (pCR2.1-TOPO vector, TOPO-TA Cloning kit and electrocompetent cells TOP 10; Invitrogen, Carlsbad, CA) and sequenced.

Processing sequencing samples

Sequences were processed using a data curation pipeline implemented in MOTHUR [93], which removed sequences from the analysis if they were less than 200nt or greater than 600nt, had a low read quality score (< 25), contained ambiguous characters, had a non-exact barcode match, or had more than 4 mismatches to the reverse primer sequences (926R). Remaining sequences were assigned to samples based on barcode matches, after which barcode and primer sequences were trimmed. Chimeric sequences were identified using the ChimeraSlayer algorithm [94], and reads were classified with the MSU RDP classifier v2.25 using the taxonomy maintained at the Ribosomal

Table 2.13: 454 barcodes and primers

barcode seq	"A" barcoded adapter for XLR system + barcode + V3–5 926R primer
CACGC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGCACGCCCGTCAATTCMTT-TRAGT
CGCAAC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGCGCAACCCGTCAATTCMTT-TRAGT
TGAAGC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGTGAAGCCCGTCAATTCMTT-TRAGT
ACTTGC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGACTTGCCCGTCAATTCMTT-TRAGT
TCACAC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGTCACACCCGTCAATTCMTT-TRAGT
CGTGAC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGCGTGACCCGTCAATTCMTT-TRAGT
ACGCGC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGACGCGCCCGTCAATTCMTT-TRAGT
CCTCTC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGCCTCTCCCGTCAATTCMTT-TRAGT
ACTCAC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGACTCACCCGTCAATTCMTT-TRAGT
AGACAC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGAGACACCCGTCAATTCMTT-TRAGT
CGACTC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGCGACTCCCGTCAATTCMTT-TRAGT
AGCTTC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGAGCTTCCCGTCAATTCMTT-TRAGT
AAGCCGC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGAAGCCGCCCGTCAATTCMTT-TRAGT
CAAGAAC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGCAAGAACCCGTCAATTCMTT-TRAGT
AGTTGGC	5' CCATCTCATCCCTGCGTGTCTCC-GACTCAGAGTTGGCCCGTCAATTCMTT-TRAGT
TATCAAC	5' CCATCTCATCCCTGCGTGTCTCC-

Database Project (RDP 10 database, version 6) [95]. After processing, the resulting sequencing depth was 2690 ± 898 (median ± median abs. deviation) reads per sample.

Supervised Learning in Microbial Ecology (SLiME)

Using a set of training data, supervised learning algorithms can be trained to classify each microbiota sample into distinct classes (eg. IBD/non-IBD) based on a defined set of features (eg. the relative abundance of each OTU). We first assigned each sequence in the data set to a taxonomical group using the RDP Naive Bayesan classifier. [96] For each sample we then calculated the relative abundance of each taxa with respect to the total number of sequences in each sample. We then trained a random forest (RF) classifier (R-project implementation [97] [85]) to assign the class (IBD or non-IBD) based on the relative sequence abundances in every t extbackslashaxa. We used ten-fold cross-validation to compute accuracy of the classifier, where training of the classification algorithm employs a random 90% of the available patients and the performance of the generated model is tested on the remaining 10% of patients. Performance of the resulting model can be described by using a receiver-operating-characteristic (ROC) curve. All plots were generated in R. [98]

Fecal calprotectin test

Calprotectin was assayed using the calprotectin ELISA kit (ALPCO, Salem, NH) and followed the manufacter testing protocol. Samples were shaken on an orbital shaker at 600 rpm. ELISA plates were read with the Varioskan (Thermo Scientific). SkanIT software (Thermo Scientific) was used to fit the standard curve using four parameter curve fitting.

Statistical analyses

Several approaches can be used to identify the features most important to the classification: a) a priori statistical tests, b) statistics intrinsic to the supervised learning algorithm or c) iterative

measures of the importance of each variable. [99] To minimize computational complexity we chose to employ an a priori statistical test. Taxa were tested for significant association with disease state by means of non-parametric Kruskal-Wallis test, which does not include an assumption of normality. Multiple p-values were then converted to q-values, by FDR adjustment [87] and a significance threshold was chosen between q-value < 0.01 or q-value < 0.05 by estimating the $\pi 0$ parameter as well as the number of false positives vs. cutoff (see [87, Storey] for details). In the case of IBD/control, CD/UC and activity classification, features individuated by Kruskal-Wallis test were largely overlapping with the list of most discriminative features obtained by iterative measures and intrinsic measures (data not shown).

Receiver operating characteristic analysis was used to evaluate the classification algorithms across a range of possible disease prevalences. Reported AUC values are median AUC values resulting from 3 repetitions of 10-fold cross validations. All calculations and plots were performed in R. [97] [98]

Chapter 3

SLiME: Synthetic Learning in Microbial Ecology

Eliseo Papa, Eric J. Alm

This chapter is presented as prepared for submission

Abstract

Complex microbial communities have recently begun to be characterized by means of high-throughput sequencing and proteomics. As a result, the burgeoning field of metagenomics and microbial ecology is generating a large number of complex high-dimensional datasets, which pose a considerable analytical challenge. In order to gain insight from these systematic efforts in the near future, it will be necessary to correlate compositional and functional data and mine it for the presence of biologically relevant patterns. Most analytical strategies which have been proposed thus far tend to be descriptive, often resulting in long lists of genetic and functional components ascribed to each community. Few strategies have been capable of highlighting important biological pattern between communities and infer the underlying molecular mechanism, largely due to the sheer size and complexity of the data involved. It is only recently that more advanced data mining algorithms have begun to be applied to the study of these complex biological datasets. Machine (or synthetic) learning algorithms, in particular, are designed to rapidly infer the discriminating characteristics and defining features of large amounts of complex structured data. The increasing abundance of large complex datasets in a number of disciplines have made machine learning the subject of intense research efforts and have rapidly advanced the capabilities of these algorithms. We and others [84] have recently demonstrated the succesful application of these algorithms to microbiota data. In order to favour further applications of these analytical techniques, we have developed a software tool, SLiME (synthetic learning in microbial ecology) and a publicly available web interface. Here we characterize its performance on benchmark datasets, illustrate potential uses and outline some of the architectural choices made. Lastly, we discuss planned features and outstanding issues in the application of machine learning to microbiota data.

3.1 Introduction

The advent of faster and more affordable sequencing technologies has allowed many innovative studies in the field of microbial ecology. Community studies of microbial diversity have been conducted for a variety of natural environments, including saltwater [100] [101] [102] [103], salt marshes [104] and lakes [105]. These studies have allowed the inference of information about community structure, number and type of ecological niches, mechanisms of bacterial communication and population-wide metabolic networks [106] [107]. It has also been possible to observe cooperativity and group behaviors in reaction to external stimuli and evolutionary pressures [2].

Efforts are now well underway to describe the microbial diversity of the human-associated microbial flora, or human microbiome [108], and its relation to human disease. [109] The Human Microbiome Project (HMP) aims to provide a map of the bacterial diversity found in association with humans. Through the use of metagenomics and whole genome sequencing of microbial flora samples from the skin, vagina, mouth, gut and nasal passages of hundreds of individuals, this project aims to establish associations between the make-up of microbial populations and health or disease. [109] [108]

The accumulation of these large datasets, has made increasingly clear that gathering insight into the microbial ecology of microbial communities will chiefly depend on the ability to analyze the data in a scalable and meaningful way. However, the sheer size of microbial sequencing data being accumulated poses a significant logistical and computational challenge to this goal. In addition, the inference of biological mechanisms or disease association is currently hindered by the lack of appropriately sophisticated analytical tools. The large number of species which can be found in an human microbiome sample, makes assembly and annotation considerably more challenging. [110] Few of the tools classically employed in bioinformatics for comparison, prediction or detection of evolutionary signatures can be directly applied to data simultaneously originated from multiple species. [111]

The application of machine learning algorithm represents an optimal solution to this dilemma [84].

Machine learning algorithms (or *supervised* learning algorithms) allow for the automatic inference of discriminating characteristics and defining features from large amounts of complex structured data. Due to the increasing abundance of large complex datasets in a number of discipline ranging from computer science to economics, machine learning techniques have been the subject of intense research in recent years.

Employing training data, supervised learning algorithms build a predictive model (*classifier*), which can be seen as a set of rules attempting to discriminate and assign unknown samples to their appropriate class. In principle, the precision of the classifier increases with the number of input data points, thus making machine learning an ideal approach in those situations where data is abundant and the underlying model and structure of the data is too complex to be known in detail at the start of the analysis.

We decided to create a tool that would allow such kind of analysis on a routine basis, by people that may lack the required bioinformatic and computational training. Our software SLiME can be freely accessed through a web application, is simple to use, gives high-quality graphic output which can be easily exported and use in publications, require minimal input from the user and minimal pre-processing of the sequencing data. We demonstrate the use of SLiME in both benchmark classification tasks and a regression task, showing that it's well suited to the different type of metadata usually part of a microbial ecology analysis. In addition, we show that SLiME is effective on a variety of benchmark data sets and metagenomic simulated data.

Sequencing technologies and targets

The type of analysis used for microbial sequencing data directly depends on the sequencing method employed to sample the bacterial community. Traditionally, sample preparation begins with DNA extraction. DNA is commonly extracted from the environment or from an human sample by means of joint solvent extraction, mechanical disruption, and treatment with buffers of varying chemical compositions depending on the exact protocol being followed [112]. Once DNA has been extracted and purified, amplification can proceed by means of polymerase chain reaction (PCR)

with a set of desired primers. Historically, bacteria have been taxonomically classified based on PCR amplification and sequence analysis of the 16S ribosomal RNA gene [113]. This gene's secondary genetic structure is characteristic and highly conserved (likely due to its central role in protein translation), encompassing both conserved and hypervariable regions [114]. Single gene analysis of this sort is obviously less precise than Multi Locus Sequence Analysis [115] [116] or whole genome analysis, techniques that taxonomically classify bacteria based on the sequence of a set of housekeeping genes or their entire genetic material, respectively. Yet, the importance of the 16S rRNA method cannot be ignored as, in many instances, it has proven sufficient in understanding whole population structures and metabolic networks [102]. Focusing on a single gene also has the advantage of requiring relatively less expensive sequencing schemes ad thus enabling large scale and/or longitudinal studies of microbial diversity. The preparation of 16S rRNA gene libraries is arguably the most delicate step in microbiome studies and it is important to consider some of the biases it introduces in the data obtained downstream. Even though 16s rRNA has shown to be quite precise in determining the presence or absence of different microbes in given environments, it is uncertain whether observed abundance of a particular DNA sequence in the sample can be taken as a measure of the original abundance of a particular microbe. Primer selection and binding characteristics, secondary structure of genomic DNA, presence of sequence repeats and the number of hypervariable regions considered, can all strongly influence the quality and sensitivity of taxonomic reconstruction and thus the precision of subsequent bioinformatic analysis to interpret these datasets.

The choice of sequencing technology that ultimately yields the sequence of interest, be it the 16S rRNA gene or a larger genetic stretch, can also significantly bias data output . Currently the most widely used systems include the traditional Sanger method, and next-generation high-throughput methods such as the Solexa GA platform (developed by Illumina) and 454 GS20 pyrosequencing platform (developed by Roche). These next-generation technologies are actively being developed with the aim of improving upon cost, output volume, read length and speed. Choice of technology needs to be considered on the basis of the ultimate experimental goal, since each system present

with its own unique set of strengths and limitations [55]. The first generation 454 GS20 platform provided an average read length of 100 nucleotides as compared to the 36 nucleotides of the Solexa GA, however due to the specifics of its underlying biochemistry these longer reads were also prone to containing unresolved hypervariable regions [117]. The traditional, and more costly, Sanger method produces high quality reads of $\sim$ 800 bases but relies on a round of cloning and suffers of low throughput [118]. Subsequent generation of 454 and Solexa platforms, as well as the ongoing development of single-molecule sequencing by Helicos and Pacific Biosciences, should provide substantial improvements to the resolution and quality of microbiome sequencing. Efforts to uniform the sampling methods are already an integral part of the HMP [109] [108].

Traditional microbiome analysis methods

Genomic analysis of microbial communities poses different challenges than the assembly of a reference genome sequence [110]. Community structure or composition has to be inferred from phylogenetic and taxonomic analysis [114]. Current approaches that employ microbiome or metagenomic data to understand structure and differences between communities rely on the arbitrary definition of operational taxonomic units (OTUs) [57]. Sequences corresponding to each sampled microbial strain are grouped into a manageable number of units on the basis of their genetic similarities. This grouping allows for the statistical comparison of related groups from two different communities, since it becomes necessary to consider only fewer operational units rather than the entire spectra of single sequences. The threshold of similarity under which different sequences are grouped together is usually selected to obtain species-level grouping, but does not take into consideration information about the underlying ecology, phenotype or evolutionary history of the microbial community.

Many different methods exist to compare two different communities subdivided in OTUs. Metrics such as P-test, Fst, Libshuff or MANOVA can estimate the statistical significance of the difference in OTU composition between two different communities [119]. P-test, Fst and UniFrac in particular tend to be the preferred choices, as they do take phylogenetic information into account

[120]. The presence or absence of certain OTUs, or their estimated abundance in the sample, is used in various ways to produce a global measure of similarity or, more fittingly, dissimilarity. Methods such as principal component analysis (PCA) can be used in a similar manner or in conjunction with these methods to visualize and quantify differences between communities. However, there exists several disadvantages to these analytical methods. Summarizing in a single value the difference between two extremely complex bacterial communities is bound to have negative effects on the ability to distinguish closely related groups. By losing resolution on the finer architecture of the bacterial groups, it becomes difficult to deconvolve the actual differences observed between communities. Measures such as P-test or Unifrac are only capable of producing a single value quantifying the relatedness between two microbial communities, but fail to produce a list of taxonomic groups (or thresholds in the abundance of certain microbial groups) capable of discriminating between samples. Although PCA determines how different two or more communities are, it does not permit identification of which microbial groups can be used to distinguish between communities. Perhaps most importantly, when employed for diagnostic purposes on human samples, global metrics have not proven sufficient to discriminate between diseased and control patients with reasonable accuracy. A single measure may not distinguish between individual to individual variations and truly meaningful biological differences between diseased and healthy patients. In one of the few clinical studies to date, clustering by PCA was shown to be relatively precise (ie. high positive predictive value) in identifying patients with inflammatory gut disease, but suffered from extremely low sensitivity (ie. low recall or a high number of false negatives) [7].

Approaches based on classification

Some of the limitations of the current analytical techniques can be overcome by implementing approaches based on learning and classification. Instead of measuring microbial diversity in communities based on their global composition and our current hypotheses about the underlying microbiology, the data itself can be used to infer the key elements which differentiate multiple microbiomes. Inferring patterns from complex data in an automated fashion should elucidate an increas-

ing number of associations that would otherwise be overlooked by clustering and visual inspection.

Machine learning is a discipline at the intersection of artificial intelligence and statistics that is concerned with this exact problem [121]. Many machine learning techniques have been developed for the automated classification of unknown samples based on the recognition and learning of complex pattern in a training dataset. In its application to microbiome data, knowledge associated with the individual microbiomes (whether the individual is diseased or healthy, for instance) is used to train a classifier capable of placing the elements of the training data set into the correct classes. When applied to new samples (a patient for which the diagnosis is uncertain, for instance) the classifier predicts with a certain accuracy which class the sampled microbiome is part of (eg. healthy or diseased). The features or underlying structure of the data individuated by machine learning algorithms provides a *local* high-resolution measure of diversity (in contrast with the *global* values produced by P-test, UniFrac, etc.) and should prove significantly more robust in accounting for individual to individual variation.

Synthetic learning algorithms require minimal user input often allowing to reduce computationally expensive pre-processing steps. Importantly, machine learning does not explicitly require knowledge of the specific domain being classified, hence its wide applicability to diverse problems. Often this translates into insights that would otherwise being overlooked, if domain specific theory was applied to the analysis of the data. Developments in other fields can be carried over to the specific domain of interest and specific classification problems can easily be shared with a wider community. However, the lack of domain specific structure can also mean that deducing mechanistic details from these techniques may be challenging. As a result of their widely applicability, many synthetic learning algorithms exist in literature and a large portion has already been applied in the biological field.

One important implication of learning algorithms is that accumulation of larger training data set should in principle improve classification accuracy and, as a consequence, knowledge of the system. Diagnostic and analysis tools based on these algorithms would therefore become increasingly precise and accurate when employed on larger datasets. Thus, by identifying the key features in the

data that distinguish diseased and healthy patient, or different ecological habitats, machine learning will permit the development of cheaper and more effective diagnostic tools, guide biological discovery and inform future sequencing efforts.

3.2 Results

Classification can accurately distinguish body sites

To test the performance of classification algorithms applied to microbiota we selected the Costello body sites dataset [122], which contains sequence data for 602 samples taken from a collection of individuals and body sites. Our algorithm of choice is the widely used randomForest algorithm [85], which has already been demonstrated to work with high accuracy across a series of microbial datasets [84]. To build a training dataset we employ the RDP naive bayes classification and avoid binning the data into OTUs, to reduce the computational complexity of the pipeline. We then count the occurence of each sample at each taxonomical level individuated by the naive bayes classifier. RandomForest is then run in weigthed mode with 10-fold cross validation and the resulting measures of variable importance as well as sample classification probability are recorded. When applied to the classification of body habitat, SLiME was capable of classifing samples on the basis of their origin from the skin very succesfully (Figure 3.1 is a sensitivity-specificity curve for the classifier), resulting in an almost perfect classifier.

SLiME regression can predict nitrate levels in bioremediation pools

The approach packaged in SLiME can also be applied with success to data with continuous labels, as often found when microbiome data is sampled along physical variables in an environmental gradient. To illustrate the capabilities of SLiME in this *regression* mode, we selected a dataset describing the microbial composition of communities found in uranium bioremediation pools [123]. Community composition as determined by sequencing was compared with various physico-chemical

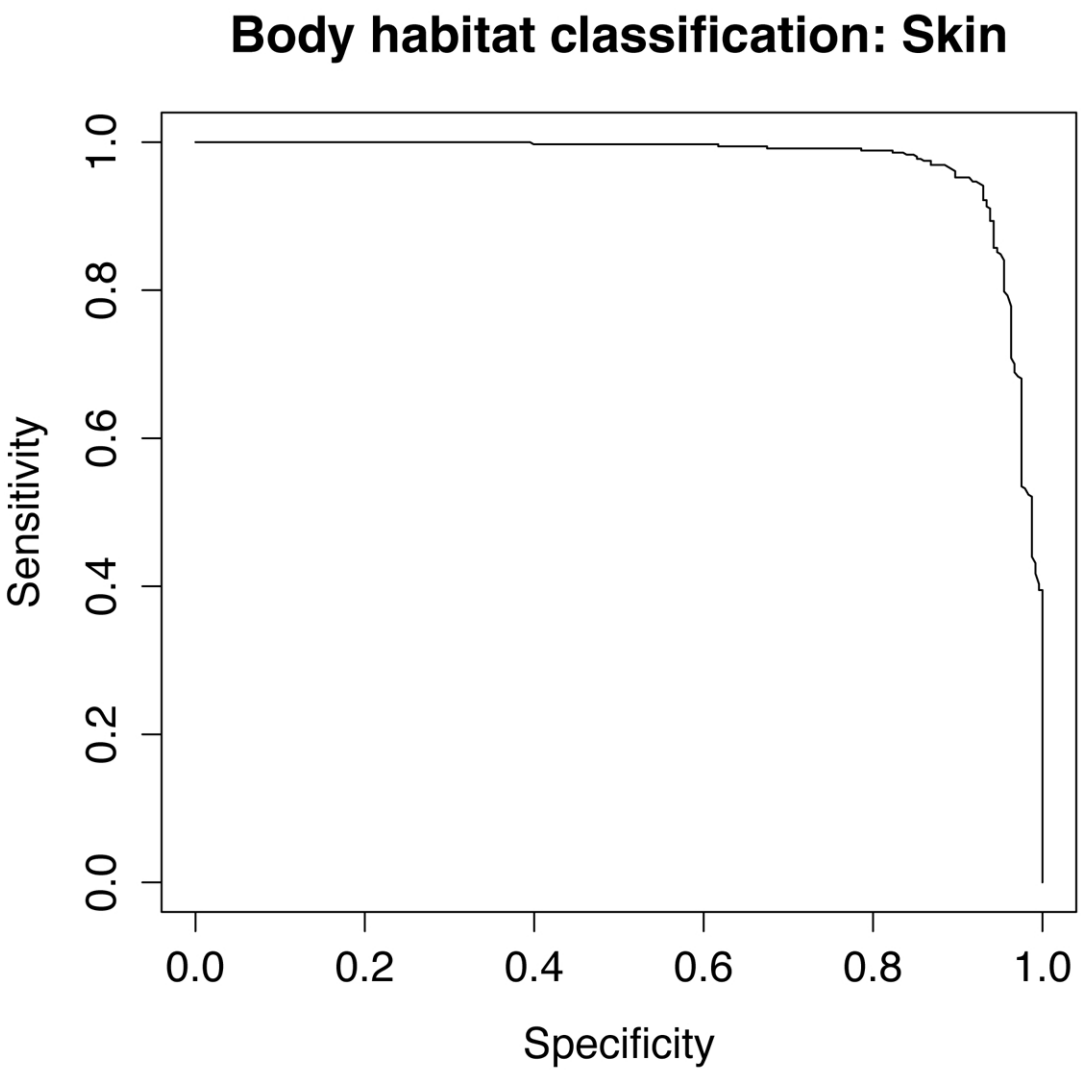

Figure 3.1: Sensitivity-specificity curve for SLiME applied to the Costello body habitat dataset. SLiME can classify whether the sample was taken from the skin of each individual with high accuracy. A perfect classifier would lie on the top right corner with 100% sensitivity and 100% specificity. A random classifier would consist of a diagonal line from top left to bottom right.

parameters to individuate potential correlation. After training, and on the basis of microbial composition alone, SLiME was able to accurately predict both nitrate NO_3 levels (Figure 3.3) and chemical oxygen demand (COD) (Figure 3.2), an indirect measure of the amount of organic compounds found in water. SLiME was precise in predicting nitrate, with the exception of two samples for which the predicted value is a gross underestimate. The predicted values for COD were also an underestimate, however the shape of the correlation is obviously linear and would allow for systematic adjustments if SLiME was routinely applied to this type of datasets.

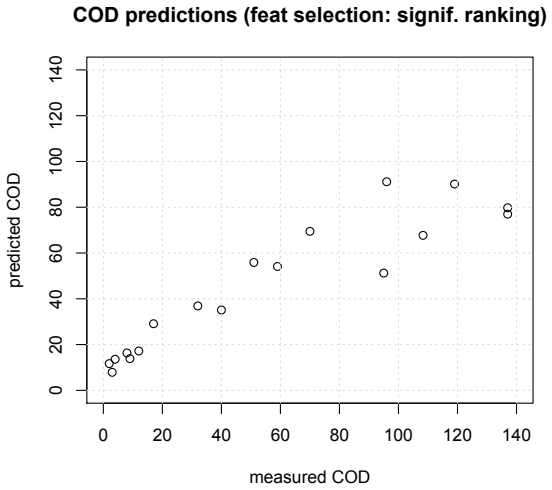

Figure 3.2: Chemical oxygen demand predictions by SVM regression. Support vector regression was employed to predict the COD value on the basis of microbial composition. The prediction measure is an underestimate of the real value, but is consistent throughout microbial samples and can therefore be systemically adjusted.

Naive assignment to taxonomic groups is sufficient for accurate classification

Considering the capability of SLiME to classify microbiota on the basis of continuous and discrete labels, we wanted to understand if the classification accuracy could be improved by different data pre-processing steps. Classically, all analysis of microbial community composition rely on the construction of tables quantifying the occurrence of each OTU in the samples. Nevertheless, the

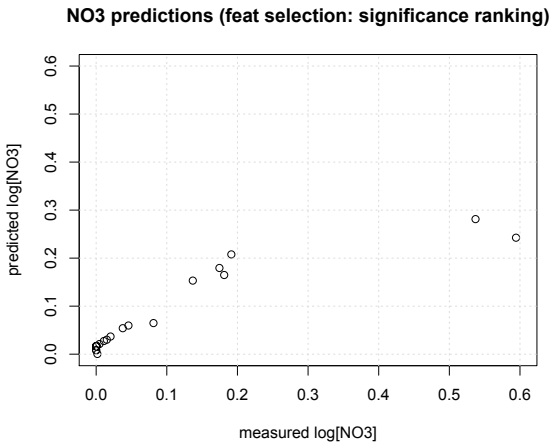

Figure 3.3: Nitrate prediction by SVM regression. Support vector regression was employed to predict the logarithm of the nitrate value on the basis of microbial composition. Notice the almost perfect correlation between predicted and actual values, with the exception of two specific samples.

construction of OTU tables can be a computationally intensive task, especially with larger dataset size. Moreover, different approaches can be used to build OTUs: selecting an identity threshold percentage or aligning the sequence to a reference phylogeny. How these two OTU deducation methods affect the final result is not known. An alternative approach to building an OTU table is to rely on the rapid approximation of taxonomical lineage, such as the one offered by the RDP naive bayes classifier. This is a quick but redundant approach, since no effort is taken to limit the occurrence of the different samples multiple times over taxonomical levels. Whether redundancy negatively or positively affects classification accuracy is poorly understood. To address some of these questions we performed a SLiME classification task at a range of OTU identity thresholds, as well as on the naive RDP classification.

While tables derived from the naive RDP classification yielded lower classification accuracy than non-redundant OTU table, the computational complexity required to compute OTU tables makes naive RDP binning a valid approach (Figure 3.4). Its validity in classification of microbiota will however depend on the specific machine learning algorithm used. RandomForest and other ensemble algorithm may be particularly robust (and indeed benefit) from the accumulation of

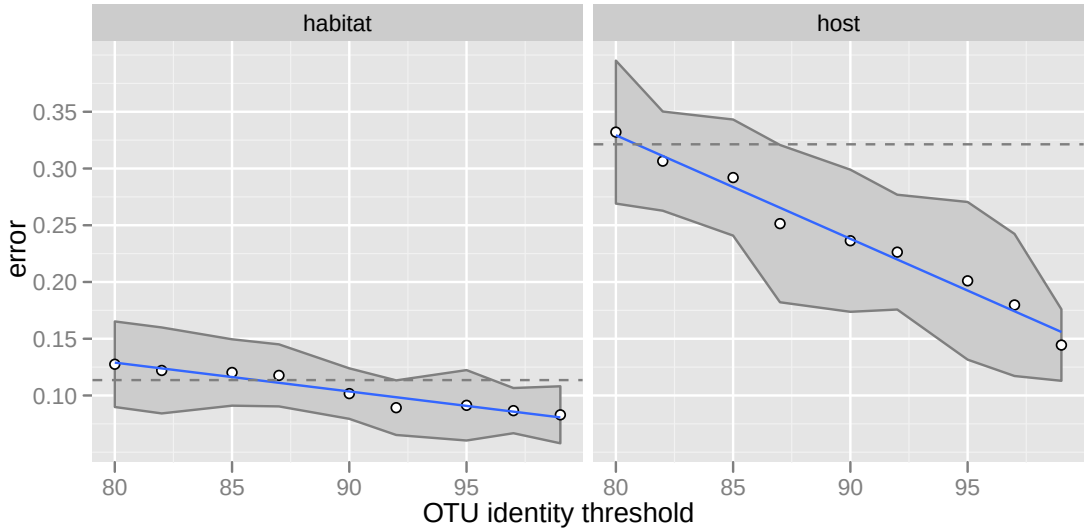

Figure 3.4: Classification error as a function of OTU identity threshold employed in the dataset construction. De novo clustering of sequences in OTU based on identity threshold may be done at any arbitrary threshold, generating a different OTU table depending on the value of this parameter. SLiME was applied to each of the resulting OTU tables and a linear fit was used to describe the trend observed. Ribbon indicates the variability of the test error (standard deviation) as estimated by 10-fold cross validation. The dashed lines indicates the performance of SLiME on tables generated by the much faster RDP binning method.

semi-redundant data and will automatically avoid problems of *overfitting* (A model which is overfitting may be really accurate on the training/test data sets, but may be harder to generalize to less similar datasets). Other algorithms which do not rely on ensemble methods may be negatively affected by the presence of multiple correlated data and indeed suffer from overfitting.

Computational complexity increases considerably with the number of OTU analyzed

We analyzed the computational complexity (and time taken) for each step in the computational pipeline employed by SLiME. While the specific implementation found in SLiME may differ from other software, similar steps are found in any machine learning pipeline deployed thus far. Specifically, we focussed our attention on those steps that occur after a compositional table is submitted by the user. Different possibilities exist: random forest can be run directly on the unfiltered table, or rarefaction and filtering can reduce the size of the table before random forest is run.

Filtering, in particular, addresses the problem that 16 rRNA sampling of a particular environment at a specific time may include spurious OTUs. High transfer from a productive environment (source) to a poorly productive environment (sink) may mean that the majority of the bacteria sampled in the later environment are only transient members of the microbiota. In the gut, lumen bacteria may be heavily influenced by the ingested bacteria and sampling may not represent the resident population of the gut. Filterint OTUs which occur only in a small fraction of samples is a one potential way to reduce the impact of this sampling bias. Rarefaction, on the other hand, addresses the differences in sequencing reads between samples that are often the result of shotgun sequencing methods. By downsampling the dataset to a minimum number of OTUs present in each sample considered, these differences are reduced. The downside of this approach is the potential loss of information caused by ignoring a large portion of sequencing data.

Figure 3.5 shows that the time taken by each classification step increases almost exponentially with the number of OTUs considered. Even though randomForest is a fast algorithm, any table

containing more than ∼500 OTUs would involve a noticeable amount of time (from above 10min to 1 hour on a typical scientific computer cluster at the time of writing) for processing. If however, rarefaction and filtering are performed before classification, the time required for classification is substantially reduced.

A tradeoff thus exist between preprocessing steps (rarefaction and filtering) and the time taken on the classification step (Figure 3.6). Rarefaction and filtering of the compositional table is also computationally burdensome and the cumulative time involved in preprocessing and classification needs to be weighed against the time taken by the classification step on the raw unfiltered dataset.

In the implementation chosen for SLiME both non-redundant OTU tables and RDP binning methods are allowed. However we opted to rely on RDP binning for the default output, as the computational costs it incurs are minimal.

Sample rarefaction and filtering does not affect classification accuracy

Since filtering and rarefaction steps may significantly reduce the size of the data, we wanted to assess if any decrease in classification accuracy could occur as a result of the filtering and rarefaction step. Promisingly, no noticeable difference exist (Figure 3.7) between SLiME classification as performed on filtered compositional tables or raw count tables. This applies when classifying different variables, across a range of OTU threshold, and also when employing RDP naive bayes binning.

Normalization does not reduce classification error

If the rarefaction step is bypassed, the difference in the number of sequences between samples can be a source of bias for the classification algorithm. To investigate whether the randomForest implementation contained in SLiME was sensitive to this bias we compared classification performance (10 repeats of 10-fold cross validations) of SLiME when fed normalized taxon abundance data or raw counts. No difference was noticed in the classification output of the two types of input data, even across different variables.

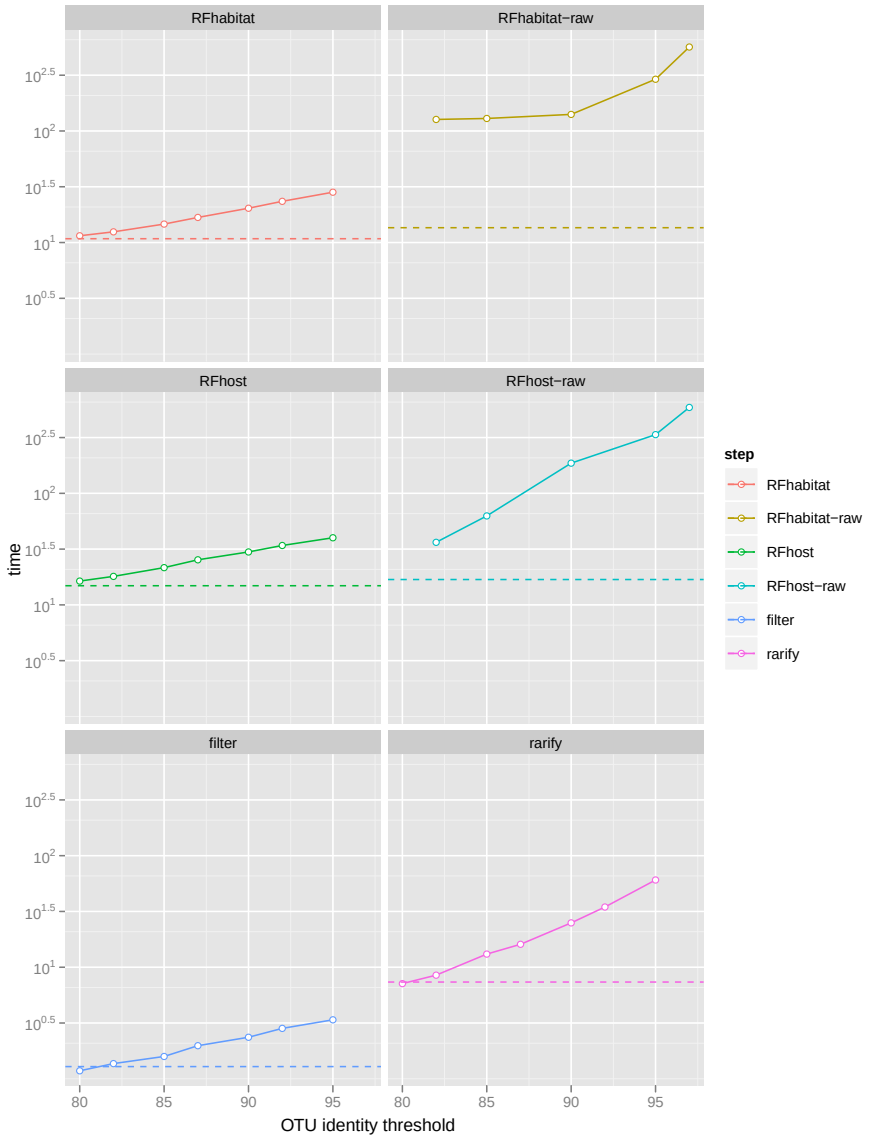

Figure 3.5: Computational time measured as a function of OTU identity threshold. Time was measured by the UNIX command `time` on the computational server beagle.darwinproject.mit.edu. Dashed line in each panel represent the time taken when fed the RDP naive binning compositional table, which involved the least computational complexity.

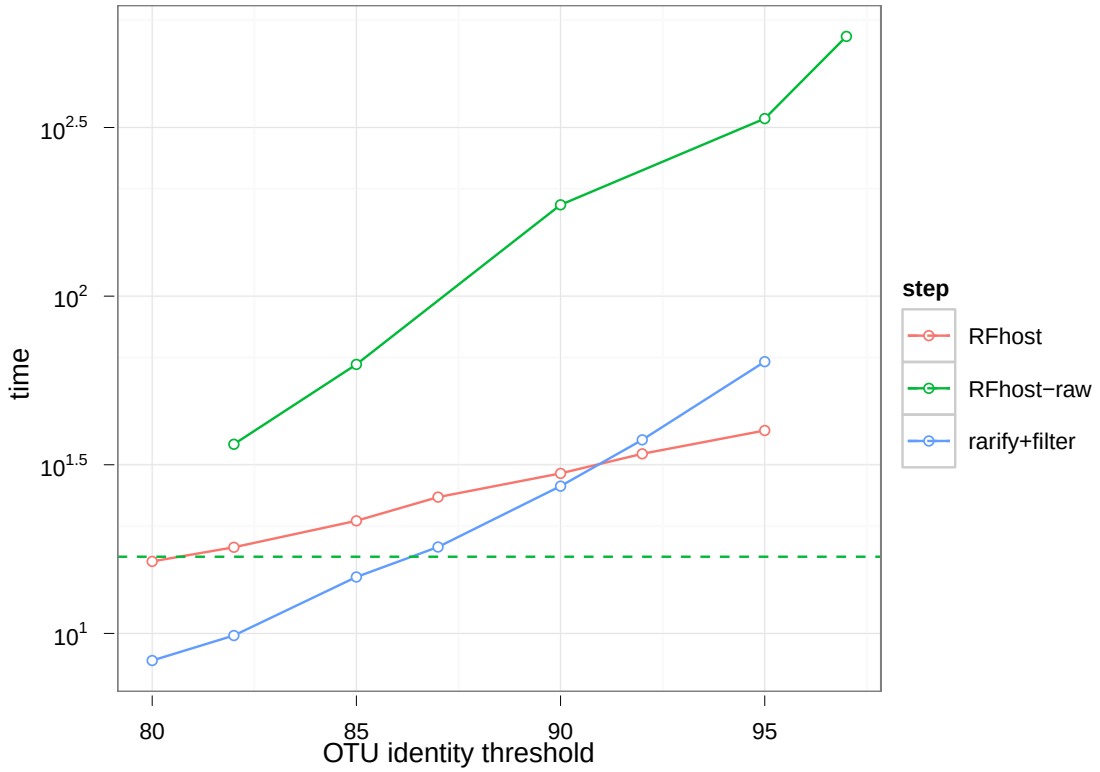

Figure 3.6: Trade-off between rarefaction, filtering and classification. Above 92% identity threshold, the preprocessing steps become the main driver of overall computational time. When the classification step is performed on raw data, the performance is substantially worse than with filtered data.

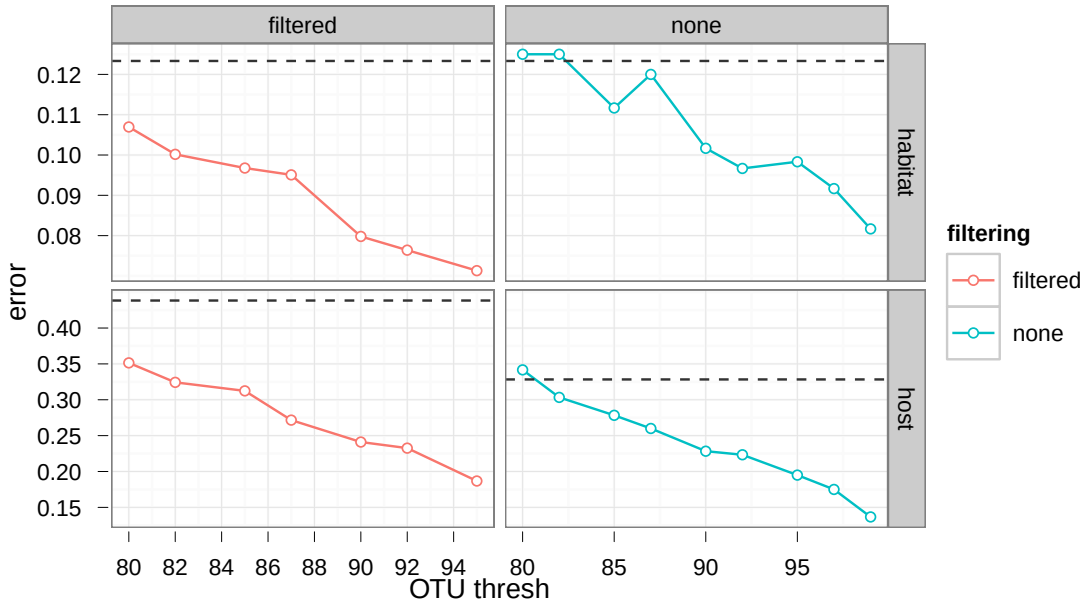

Figure 3.7: Error as a function of OTU identity threshold and its dependence on filtering. OOB error for SLiME across a variety of OTU identity threshold was performed on the habitat benchmark and the host benchmark. No substantial difference appears between filtered abundance data and raw read counts. Dashed line show the performance obtained by the RDP-binned table.

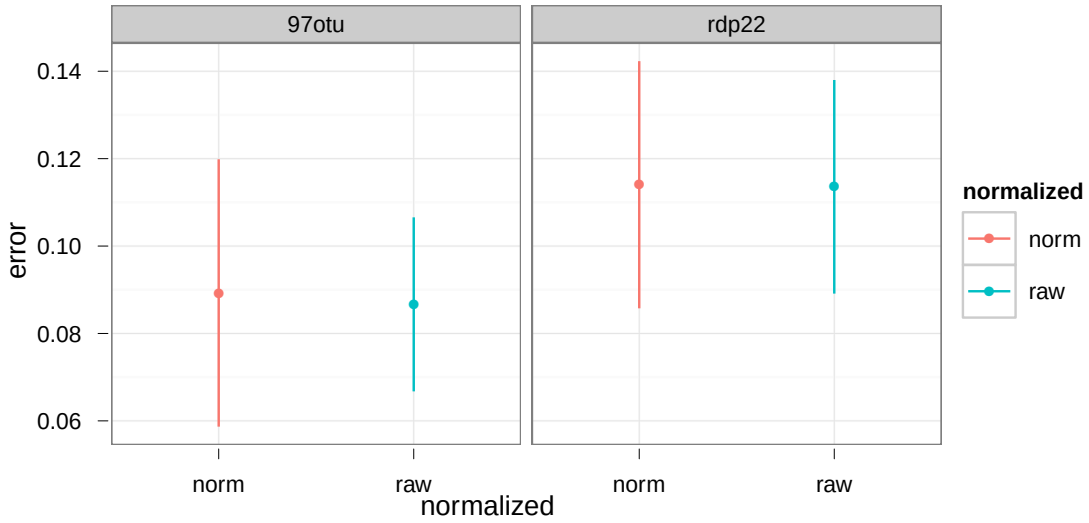

Figure 3.8: Normalization does not affect performance. Normalizing the compositional table with respect to the total number of reads per each sample does not alter the outcome in any significant manner.

Classification accuracy improves with the number of samples

One attractive property of machine learning algorithms is that in principle their accuracy will improve as more training cases are added to the datasets. The study of this correlation between model accuracy and sample size remains largely empirical, but this trend has nonetheless been demonstrated with many algorithms, both in balanced and unbalanced data sets [124]. To test this property we formed a series of random subsets from the IBD Children's dataset and fed them to SLiME, measuring the area under the resulting ROC curve. The data was then fitted with a local regression (LOESS) and plotted in Figure 3.9. Confirming expectations, a non-linear trend upwards was observable and, at least for the number of samples available, it was not possible to determine at which level would it reach an asymptote.

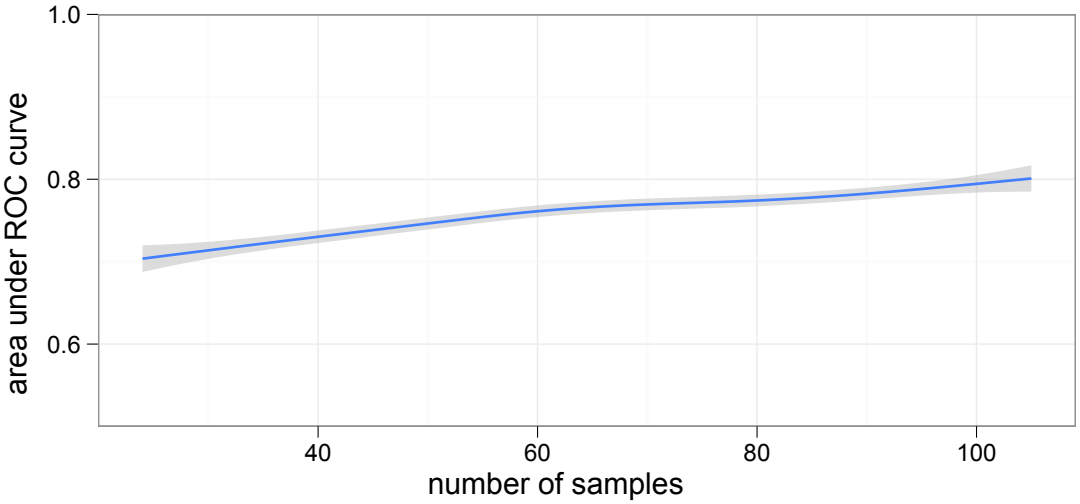

Figure 3.9: Area under the ROC curve increases as more training samples are added. The area under the ROC curve for the classification of patients in IBD or healthy was measured while rarefaction of the number of samples included was performed.

Simple heuristics for phylogenetic data incorporation do not improve classification accuracy

It is well established that microorganism distribution across habitats is correlated with their phylogeny, such that different habitats can be distinguished on the basis of beta diversity. [125] Comparing 16s rRNA surveys and other phylogenetically informative marker genes [126] showed that this correlation is relatively stable and that habitat preferences do not change significantly in evolutionary time. However, this is not to say that phylogeny can completely describe the distribution of microorganisms in their habitat and that other phenomenon (such as horizontal gene transfer) take place. A detailed prediction of the microbe in each environmental niche need to take into account other elements beside phylogenetic composition, although this serves as a good first approximation.

Considering the reliability of phylogenetic information in predicting microbial distribution, it is reasonable to assume that including phylogenetic structure in the data analyzed by machine

learning would improve the performance of the prediction. Measures such as unifrac have been shown to cluster microbial compositional data better than non-phylogenetic measures such as bray-curtis [127]. On the other hand, because different datasets would require building separate phylogenies, employing phylogenetic information in the classification task will prevent simple comparison of datasets, unless a common reference phylogeny is agreed upon and employed. To test the validity of phylogenetic information in aiding classification tasks we attempted to build training tables employing on a simple phylogenetic heuristic (see Figure 3.10) and performed classification tasks on the Children's Hospital IBD dataset previously described.

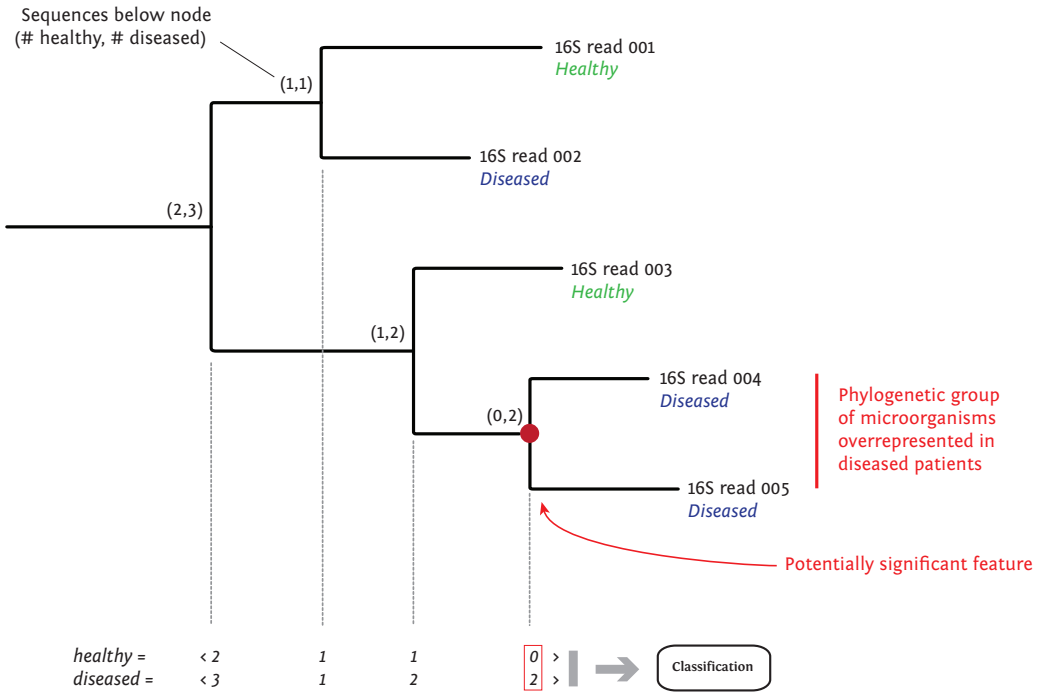

Figure 3.10: Phylogenetic information heuristic. Training tables which incorporated phylogenetic information were constructed by assembling a tree from the submitted sequences and traversing the tree from leaf to root. At each node, the number of leafs present in each sample was counted, weighted by either distance to ancestors or node depth, and further propagated recursively.

We used both SLiME and other algorithms such as hierarchical decision trees ([**?**]) and measured the classification accuracy as the diagonal of the resulting confusion matrix. Average

accuracy across a variety of condition was 80%, which was lower than observed with SLiME.

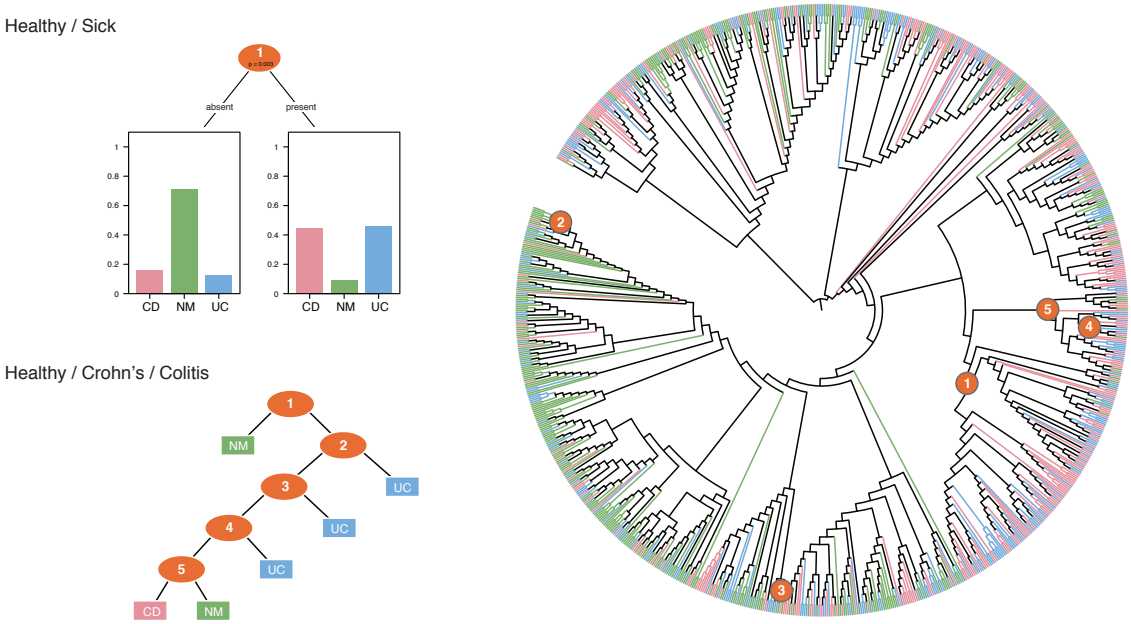

Figure 3.11: Hierarchical decision tree outlining the classification of a patient as normal, crohn's or colitis, depending on whether sequences are present at the given nodes in the phylogenetic tree. Average accuracy is 80%. Decision tree nodes are colored with respect to the hierarchical level. Tree branches are colored based on diagnosis. Bacterial groups in a normal patient are colored green; magenta for Crohn's samples and cyan for colitis samples.

To address whether the observed decrease in performance was due to the classification algorithm, we also tested the same heuristic with support vector machines (SVM). We found (Figure 3.12) an improvement in average overall accuracy to 85% which was also reflected in multiclass classification, yielding an average of 72% accuracy.

Thus, while in principle phylogenetic information should augment the data available to machine learning algorithm and thus allow an improvement in performance, when we include phylogenetic distance by means of a simple heuristic, we found minimal change in performance. The improvement was not significant enough to justify the increase in computational complexity and the inability to compare different data sets quickly. Obviously more work is necessary to individ-

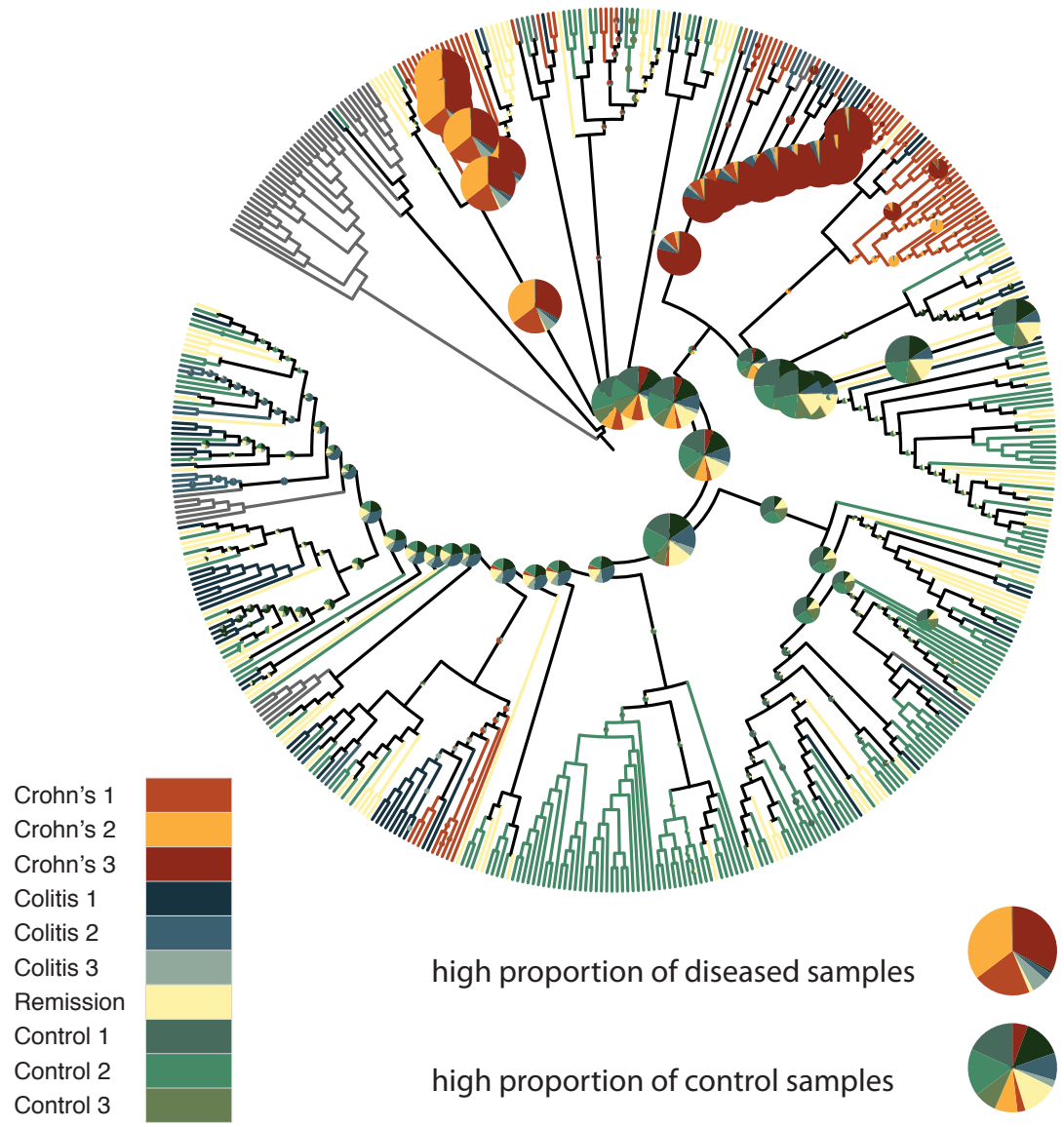

Figure 3.12: Phylogenetic feature selection for SVM classification. Nodes in the phylogenetic tree with a disproportionate ($p < 0.05$) number of disease-associated or health-associated sequences are chosen as features for SVM. Cross-validated classification leads to 85% average accuracy in distinguishing healthy and diseased patients. Distinction in the three classes (control, colitis, Crohn's) is 72% accurate. At each node the proportion of sequences for each patient in the subtree is indicated. Branches are colored based on the corresponding patient diagnosis (green for control, blu for colitis, yellow for remission and red for Crohn's)

uate other means of including phylogenetic information.

One of the main disadvantage of incorporating phylogenetic information is that comparing multiple data sets requires the construction of a common tree. Another approach, of which the RDP naive bayesian classifier is a basic example, is the assignment of reads onto a reference phylogeny. Given an existing phylogeny, it is computationally far more tractable to assign reads by similarity to the different phylogenetic levels, without recomputing the base phylogeny. Software tools such as RaXML have been adapted and used for this purpose [128]. However, this requires the reference phylogeny to be of high standard and completeness, enough to be agreed upon by researchers in different fields and who plan to align very disparate data sets. This remains largely unachievable at present, given the varying quality of sequencing data across the Tree of Life. Moreover, a reference phylogeny that does not incorporate the data at hand in its construction may be significantly less precise, particularly at the taxonomical levels included in the analysis.

SLiME is a web app which offers interactive visualizations

While the functionalities provided by the SLiME package can be accessed from the command line under any major UNIX operative system, to further encourage adoption of machine learning analysis in the field of metagenomics, I have developed a web interface to SLiME's functionality. The application is built on a `python` stack, and at its basis lies a queueing system to which the web process passes commands and data files to be processed. The user interfaces with the web process through a user interface designed with the aid of `Bootstrap`. The worker process performs the classification and relies on the `scikit.learn` machine learning library and `matplotlib` library to output the data. Further interactive visualizations via the `D3.js` library is planned in future versions. Other planned features include the ability to choose different machine learning algorithms for classification and the online optimization of learning parameters for the random forest algorithm.

The web interface allows any user to submit either sequence data or tabular data (default) and obtain a job id upon submission. The job id is then used to poll the worker queue on the computa-

tional server and retrieve the output of the classifier. The user can then explore the importance of each variable and the probability the model assign to each label directly through a web browser.

3.3 Discussion

We demonstrate the use of SLiME in both benchmark classification tasks and a regression task, showing that it's well suited to different type of metadata which could be part of a microbial ecology analysis. To date, there are few similar applications of machine learning to microbiota. Knights et al [84] have advocated for the use of machine learning techniques applied to microbiome data and prepared benchmark data [84] to test new algorithms against. They have also showed the application of these techniques to detect sequence samples mislabelings [129] and to track the source of microbial contamination [130]. Another package (MetaDistance) recently developed addresses the sparcity typical of otu vs. samples dataset and implements distance-based learning to classify multiple labels [131]. The R package *pensim* was developed to facilitate the application of penalized regression techniques to metagenomic data, providing facilities to scale data and obviate to overfitting artefacts. Exclusively generative models have also been adapted to work with microbial composition data and metagenomics data. A variation on Dirichlet's mixture named Dirichlet Multinomial Mixtures was shown as it applies to metagenomic data from the obese/lean benchmark dataset.

These positive results demonstrate the applicability of machine learning to microbial sequencing data which I believe will have a pivotal role in the discovery of mechanisms and functions of the commensal microbiota. A potential approach to foster the power and applicability of predictive models to microbiome analysis is the launch of machine learning competitions with metagenomics data [132]. Crowd-sourcing has recently shown great potential in improving the accuracy of models for a variety of predictive task and may confer SLiME or other equivalent methods the precision necessary to be employed in the clinical and research setting.

Application of a machine learning approach to the analysis of microbial sequencing data will

also help shape future study design and data collection. Since the HMP is a communal effort amongst numerous institutes and labs all over the world, sample preparations, amplifications and sequencing approaches being used vary significantly across participants. Any tool employed to analyze the resulting data should not only correct for these potential biases, but should deliver quantitative information to aid method development. Classification - instead of the classical global quantification of diversity - can produce such information, allowing for the evaluation of the impact of the various technological parameters (sequencing platform, read length, variation between centres,etc.) and advising on the best possible strategy for any future HMP study.

Given that the potential yield of the HMP is immense, new approaches for output standardization and analysis are absolutely essential to extracting the full significance of what these data can teach us about health, disease and microbial evolution. Machine learning and the development of sophisticated learning algorithms represent a powerful and promising approach to tackle these challenges.

3.4 Methods

Machine learning algorithms

Support vector machines (SVM) is an algorithm capable of dealing with high-dimensional input data [133] [134]. Input is mapped onto a multi-dimensional space and SVM will attempts to find the separating hyperplane which leaves the largest possible margins between the two set of points in this multidimensional space. Multi-way classification using SVM can be done by partitioning the dataset and running the algorithm for each one-against-all combination of classification variable. All input variables are taken into consideration and the performance is determined by those samples closest to the generated hyperplane (support vectors). Albeit extremely powerful and rapid, this algorithm does not allow for rapid translation of the classifier in terms of discriminating microbial populations and can thus be less intuitive.

Ensemble methods To further improve classification accuracy and robustness, machine learning algorithms often combine multiple different models [135]. Such methods are called ensemble methods because they aggregate the performance of many classifiers to boost the overall classification accuracy. Boosting and bagging [136] for instance, sum the performance of many weak classifiers to obtain a strong classifier. **Random forests** [85] [137] aggregate classification trees inferred on random bootstraps of the entire sample population, take into account the result of each tree and determine the final result by combining (usually by a majority vote) these individual results. The resulting predictive model generally performs comparably to regression trees, but is less prone to overfitting. Specific weighting schemes can be employed to account for highly unbalanced datasets (data in which one class of sample constitutes a minor portion of the entire dataset, as in the detection of rare disease) [124]. The algorithm also estimates error, by applying the regression tree obtained at each bootstrap iteration to the data excluded from the bootstrap. The resulting *out-of-bag* error estimate can be used to quickly compare classification performance on the training data set and does not require additional computational steps. Lastly, random forests generate a measure of variable importance. The difference in out-of-bag classification error is measured iteratively excluding each variable from the training set, and the variables are ranked according to their impact on the final error across all bootstrapped trees.

Part III

Tools for the study of the immune system

The immune system is a highly complex network of cells that interact with each other dynamically across time and space. Traditional immunological assays are simply not capable of fully and comprehensively characterize the immune response due to their inherent limitations. Biochemical assays cannot provide the required level of resolution and high throughput necessary to form a complete picture of the immune network at any given point of time. In addition, these assays often rely on isolating immune cells from their microenvironment and exposing them to stimuli very different from the ones which they would be exposed to under physiological conditions.

Recently, insight into the etiology of complex autoimmune diseases has underscored the need of a more integrative approach (or *systems* approach) to immunity. Autoimmune diseases (IBD, diabetes, etc.) have been shown to be multifactorial, in that no single trigger is responsible for the observed phenotype. Abnormalities in the communication between immune cells, as well as a variety of noxious stimuli, all contribute to disease manifestations. Only an overall — and physiologically relevant — view of the immune response to a particular condition will allow for the characterization and prediction of the course of a disease will take.

To obtain a systems view of immunity, novel immunological assays are necessary. Tools capable of generating multivariate data with high-throughput and high resolution would allow for the generation of highly detailed snapshot of the immune system. These snapshot, in turn, could be used to construct predictive models of immunity. Assays capable of mimicking physiological stimuli and microenvironments would ensure the physiological relevance of any predictive model.

Chapter 4

Mapping rapid surface dynamics in naïve B cells exposed to subsaturating doses of ligand

Eliseo Papa, J. Christopher Love, You Me Kim, Annemarthe G. van der Veen, Hidde L. Ploegh

This chapter is presented as prepared for submission to [ISME?]

Abstract

The capture of antigen on surface-disposed receptors of B cells is critical to inducing a humoral immune response. The binding events cause 1) rearrangement of receptors into chlosterol-rich microdomains, 2) initiation of phosphorylation-mediated signaling pathways, and 3) internalization of bound antigen for proteolysis and loading onto major histocompatibility (MHC) class II complexes. These processes have been studied by biochemical and imaging methods, but these are insufficient to describe early dynamics in a quantitative manner. Here we show how a microfluidic device was used to apply sequential doses of cross-linking ligands to primary naïve B cells from mice, allowing to track the early dynamic of the labeled B cell receptors by live cell confocal microscopy. We find that upon stimulation with ligand, B cell receptors (BCRs) redistribute slowly (75.3 s), but a second exposure induces a more rapid rearrangement (35.3 s) of a distinct set of receptors. We also show that the rearrangement of surface-disposed class II MHC requires BCR ligation but not BCR microcluster formation. These results suggest that B cell respond to antigen in a manner proportional to the total amount of antigen acquired at the surface of the cell. The approach described should allow further studies on the distinct roles played by surface-bound receptors in early signaling and subsequent internalization events.

4.1 Introduction

Ligation of B cell receptors (BCRs), present on the surface of naïve B cells, triggers at least two significant events. First, initiation of a signaling cascade prepares the B cell for proliferation and differentiation into immunoglobulin (Ig)-secreting plasma cells [138] [139]. Second, internalization of the bound antigen allows its proteolysis within lysosomal compartments, and the immunogenic peptides generated associate with class II major histocompatibility molecules (MHC) [140] [141]. Presentation of internalized antigen via the class II MHC pathway allows engagement of antigen-specific CD4+ helper T cells, which then license the B cell to differentiate into memory cells or antibody-secreting plasma cells.

The affinity and the valency of the interaction between a ligand and a BCR determine how the cognate pairs are reorganized on the surface as well as internalized. Observations of both primary B cells (and immortalized lines of B cells) have shown that continuous exposure of cells to multivalent antigens leads to cross-linking of two or more BCRs at the surface of the B cell [142]. This process induces the translocation of BCRs into cholesterol-rich, detergent-insoluble microdomains (so-called lipid rafts) [143] where the BCRs have been thought to act as the primary sites for signal transduction via tyrosine phosphorylation [144] [145]. BCRs located in these microdomains recruit kinases, including Lyn, Syk, Btk, involved in the signaling cascade, and gain additional stabilization by attachment to the cytoskeleton and possibly by recruitment of other adaptor proteins [138] [146]. Monovalent antigen can also initiate both signal transduction via tyrosine phosphorylation and internalization of the BCR-antigen complex, but does so with delayed kinetics relative to multivalent antigens [142]. B cells exposed to monovalent antigen are also poor at soliciting help from appropriate CD4+ T cells.

Previous studies that have used biochemical and imaging methods to describe the dynamic translocation of BCRs on the plasma membrane have relied on applying cross-linking ligands under conditions that favor the equilibration of the ligand-receptor reaction (e.g., low temperatures at 4°C for 20–60 min), or that expose cells to a continuous source of ligand at 37°C [142] [147] [148] [149] [150].

Physiologically, however, it is likely that B cells in secondary lymphoid tissues only experience soluble antigen transiently, and in limited amounts, as it is delivered via afferent lymphatics.

Here we present new observations of the early reorganization and aggregation of both BCRs and major histocompatibility complex (MHC) class II molecules on individual primary naïve B cells that were exposed to sequential, subsaturating dosages of a cross-linking anti-Ig. To apply metered quantities of the ligand to live cells, we designed a microfluidic device for directing short "pulses" (∼20–30 s) of ligands or other molecules to the cells immobilized inside the device. This system made it possible to apply two different stimuli sequentially in time while monitoring the naïve cells by live cell imaging. We examined the effects of a single, subsaturating pulse of anti-IgM on the redistribution of both BCRs and class II MHC at the surface of naïve B cells, and the subsequent redistribution of a second, temporally-distinct population of BCRs. Quantitative image analysis of individual cells subjected to these "pulse-chase" experiments indicated naïve cells respond slowly to form aggregated patches of membrane-bound BCRs upon initial cross-linking, but subsequent exposure to cross-linking ligand induces rapid aggregation. Furthermore, the characteristic time for aggregation of MHC class II on the surface is faster than that for BCRs, suggesting that formation of BCRs microcluster after ligation is not necessary for inducing the events leading to internalization of class II MHC.

4.2 Results

Microfluidic delivery of metered quantities of ligand

We designed and fabricated a microfluidic device by soft lithography that used changes in hydrodynamic pressure to deliver defined amounts of ligands (cross-linking antibodies) to immobilized B cells. The poly(dimethylsiloxane) device comprised three parallel channels with widths of 150, 100, and 150 μm connected by a single perpendicular channel (Figure 4.1). Each channel had an independent input, but all three shared a common reservoir for waste. The relative hydrodynamic

pressure applied to each of the three input streams determined the partitioning of the streams inside the system. For routine observation of the cells, culture media was delivered through the central channel of the device; the solutions of ligands (e.g., anti-mouse IgM) were kept at rates of flow such that they remained in the two outer channels, separate from the central channel. A "pulse" of ligand to the cells was applied by adjusting the pressures at the three inlets so that one or the other stream of ligands replaced the stream of media in the central channel (Figure 4.2).

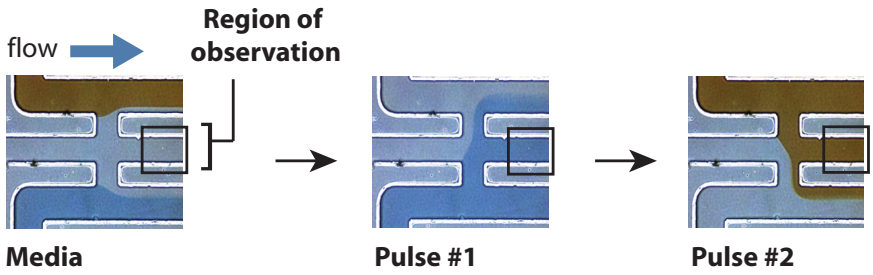

Figure 4.1: "Pulse-chase" labeling of B cells by microfluidics. Optical micrographs of the fluid flows through an example of the microfluidic device. Here, the independent flows, moving from left to right, contain PBS and two colored dyes for clarity and visualization of the design used. Varying the hydrodynamic pressure applied at the inlets enabled three possible states of the system as shown. A pulse of ligand was achieved by switching between two of the states for a defined duration. The black box indicates the location where immobilized cells were observed.

The simple design of this microfluidic system allowed control over the duration of the pulses, the amount of ligand to which the cells are exposed, and the time between two sequential pulses without the need of actuated valves. We characterized the shape of the actual pulses empirically to determine the optimal parameters for applying subsaturating dosages of cross-linking ligands to immobilized cells. To measure how the shape and duration of the pulse varied from the applied pulse (square pulse), we measured the fluorescent intensity of transferrin (Tfn, labeled with either Alexa–568 or Alexa–647) over the duration of an applied pulse (Figure 4.3). The shape of the applied pulse was somewhat different than that for an ideal applied pulse. This observation indicated that the total amount of ligand delivered through the central channel was, therefore, less than the theoretical amount of ligand that could be delivered. The observed difference was most likely due

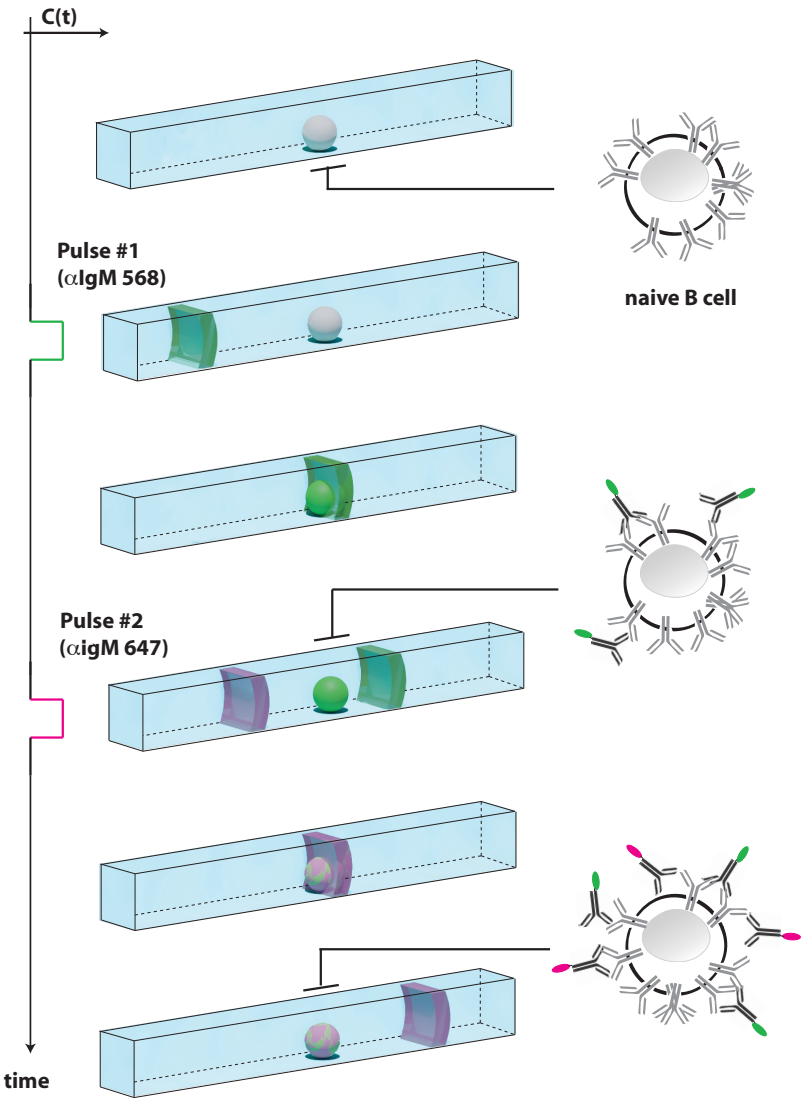

Figure 4.2: Schematic illustration of the experimental design inside a microchannel. B cells immobilized on the bottom of a microchannel experience two temporally-separate pulses of cross-linking ligands (green and red).

to the delayed mechanical response of the pumps used, and finite time required to stabilize the flow in the device. For extremely low concentrations of the fluorescent ligand, the non-linearity intrinsic in fluorescence measurements may also contribute to the delayed response time observed.

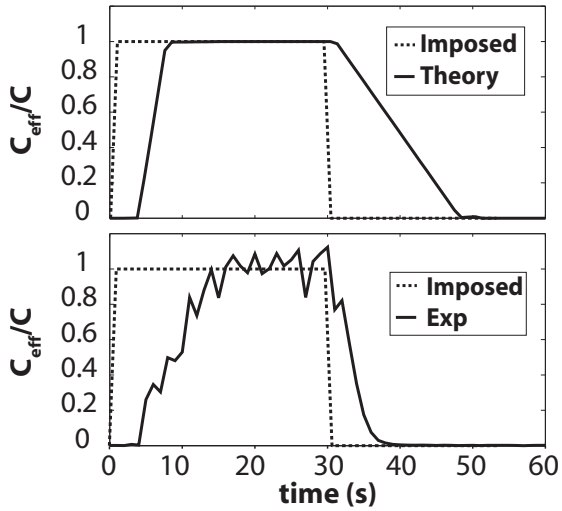

Figure 4.3: Characterization of single pulses and sequential pulses. Graphs of the expected and measured shapes of a pulse in a microchannel when a square pulse of 30 s is imposed. The quasi-periodic fluctuations observed reflect the mechanical noise of the syringe pump.

The width of a measured pulse was most similar to that of an applied pulse when its duration was less than 10 s ([1]). The effective concentration of the ligand at the surface of the cell, however, only approached that of the bulk solution for pulses lasting more than 20 s (Figure 4.5). For this reason, we chose to use pulses of 30 s in duration for the experiments described below to ensure consistent and reproducible application of ligand. It was possible to apply these pulses sequentially 1 s apart without significant overlap (Figure 4.6).

[1] uch-fig2b

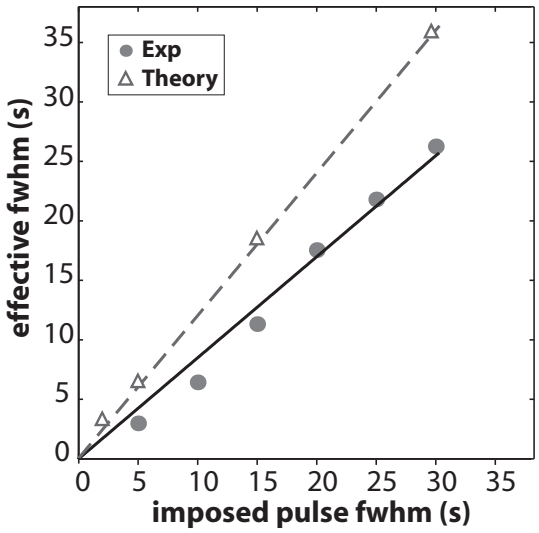

Figure 4.4: A graph of the theoretical and measured widths of single pulses as a function of duration of the imposed pulse.

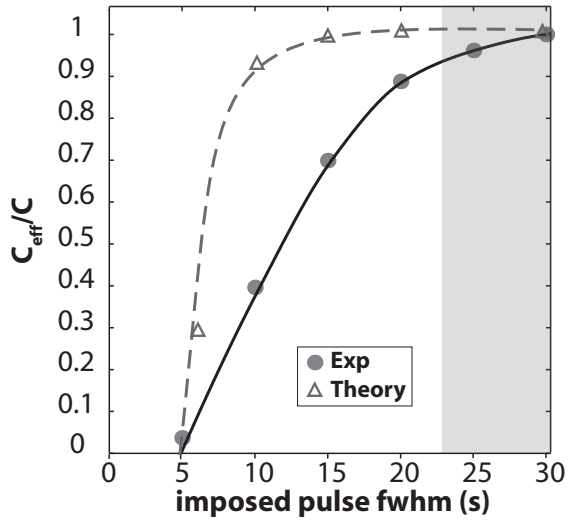

Figure 4.5: A graph of the normalized effective concentration of ligand at the surface of the channel as a function of the duration of the applied pulse. The grey shading indicates the optimal range of times required for the concentration of applied ligand to approximate that of the bulk reservoir.

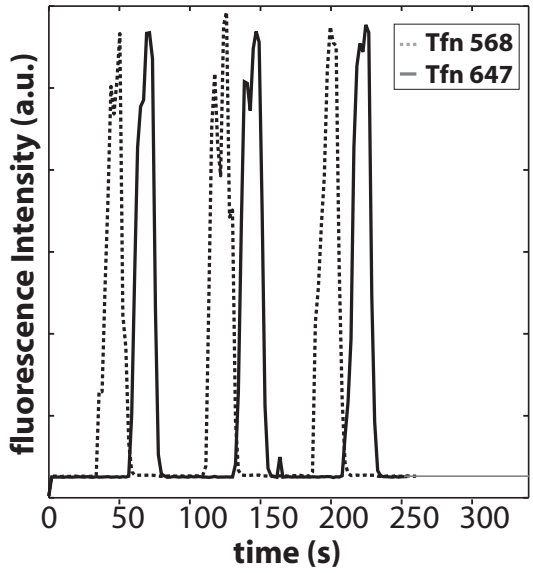

Figure 4.6: A plot of consecutive pulses of two different ligands (Tfn-Alexa 568 and Tfn-Alexa647). The imposed pulses were 30 s in length, separated by 1 s.

Limited stimulation of B cells with cross-linking ligand induces microclustering of BCRs

To examine the dynamics of BCRs on splenic B cells exposed to subsaturating quantities of a cross-linking ligand, we deposited naïve B cells purified from a transgenic mouse expressing class II MHC-eGFP into a microchannel coated with poly-L-lysine, and applied a 30 s pulse of anti-IgM-Alexa 568 (10 µg/mL) to the cells. We estimated that approximately 1–10% of the surface-bound receptors were labeled by a single pulse, by assuming a highly transient reversible receptor-ligand reaction at the channel surface (see section 4.4). Rather than employing classic equilibrium solutions, we estimated maximum receptor occupancy by evaluating the transient solution at the point in which it reaches its maximum value (here at the end of the pulse; see supplementary materials). Such low receptor occupancy suggests a regime in which unoccupied receptors are largely abundant with respect to occupied labeled receptors, thus implying that any subsequent pulse will preferentially ligate to unlabelled receptors and thus label distinct surface populations.

The population and history of cells purified from the mouse are heterogeneous by nature. We, therefore, limited our observations to cells in which the class II MHC-eGFP signal was primarily located at the plasma membrane—a phenotype consistent with naïve B cells. To maximize the time for observing the signal from the limited quantity of bound ligand, we imaged the cells at three roughly equatorial sections by confocal microscopy at intervals of 20–30s. Images of entire cells taken by confocal microscopy indicated that there was no particular spatial bias to the distribution of receptors on the surface, and that imaging an equatorial plane was sufficient for measuring the characteristic time required for reorganization of BCRs into aggregated patches on the plasma membrane (data not shown).

The population of labeled BCRs was uniformly distributed at the membrane immediately following an applied pulse of anti-IgM—that is, the measured fluorescent intensity observed on the surface of the cell was highly uniform (Figure 4.7). With time, the labeled receptors formed microclusters on the surface that were both larger and more intense than those observed immediately after the pulse. These patches likely represent growing aggregates of BCR-containing lipid rafts (also called "microclusters") because the postulated size of individual rafts (10–100 nm) is smaller than the optical resolution afforded by our microscope (~120–150 nm). Punctate structures containing ligand also appeared inside the cell 2–4 min after the applied pulse. This result is similar to the average timeframe for internalization of labeled BCRs determined by flow cytometry and other microscopy experiments. [142] [147] [148]

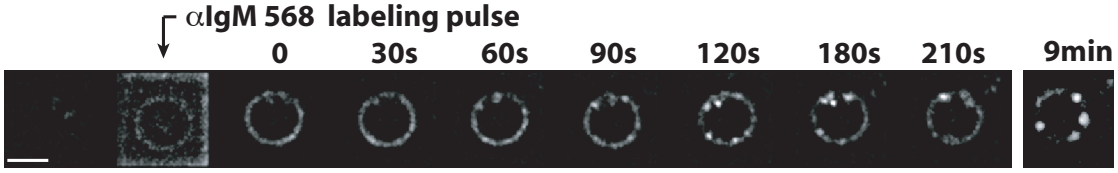

Figure 4.7: Dynamic rearrangement and internalization of labeled BCRs. Series of fluorescence micrographs of a naïve B cell before, during, and after exposure to a pulse of αIgM-Alexa 568 (10 μg/mL) for 30 s. The time elapsed after the pulse is indicated. Note the punctate structures inside the plasma membrane at 120 s. The scale bar is 5 μm.

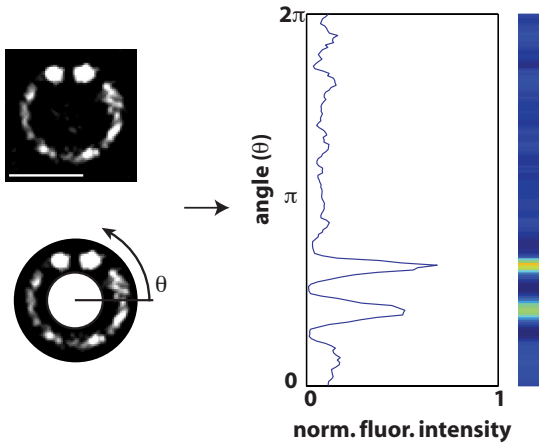

Figure 4.8: Outline of method for conversion of fluorescent intensity around the cell membrane from a single confocal micrograph to a linear density plot. The scale bar is 5 μm.

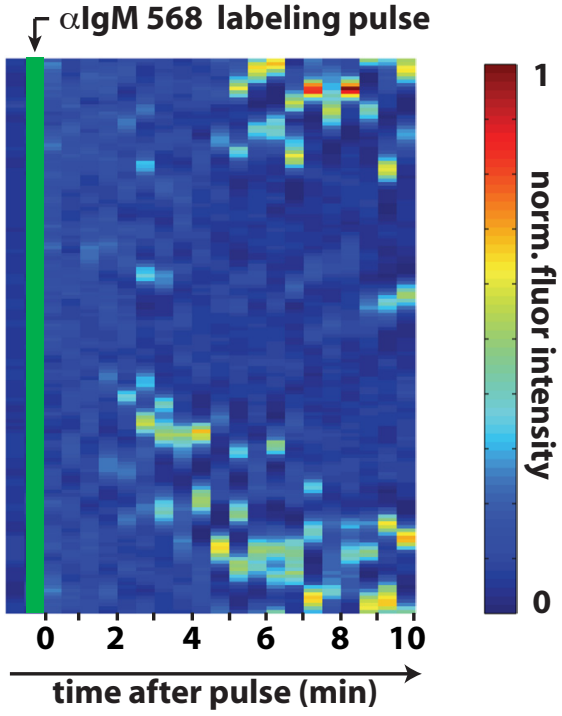

Figure 4.9: Kymograph depicting the behavior of labeled BCRs as a function of time and position on the cell membrane. The green bar indicates the time and duration of the applied pulse of αIgM-Alexa 568. Scale bar is 0.4 radians.

To further analyze the dynamic redistribution of receptors, we mapped the surface distribution of the fluorescent intensity observed for a cell as a function of time (Figure 4.9). Assuming circular symmetry in the two-dimensional optical slices, we plotted the measured fluorescence intensity in intervals of ~0.04 radians (for a typical B cell, arcs of ~150 nm in length) around the circumference of the cell within a ring of finite thickness [151]. The thickness of the ring was set to exclude punctate structures clearly separated from the plasma membrane. Visualizing the data for individual cells as a function of time showed that the number, sizes, and spatial distribution of the microclusters of BCR changed dynamically following a pulse of anti-IgM. Although some of the variations may have resulted from patches moving in and out of the selected confocal plane and internalization of some of the ligated BCRs occurred within a few minutes, the data suggest that, on the whole, aggregated domains containing cross-linked BCRs remain present and mobile on the surface for longer than 10 min.

When pulses of ligand were applied to cells, surface-disposed BCRs organized into intensely fluorescent structures of various sizes at the plasma membrane over time, but we rarely observed stable polarized, cap-like structures of BCR, even after 30 min of observation. Continuous exposure to ligand for 20 min or more in the microchannels was required to induce cap-like structures that did not change dynamically in ~60% of cells observed (data not shown). These data confirm that the size of aggregated patches of BCR, and the resulting polarization of the cell, depends to a large extent on the total number of cross-linked receptors, and suggest that considerable cross-linking of surface-disposed BCRs is required to induce cap-like structures.

Rate for reorganization of BCR increases with additional ligation

The ability to label a temporally distinct population of BCRs with a second pulse anti-IgM made it possible to probe how unperturbed BCRs redistributed on the surface of the cell following the initial ligation. We applied two sequential pulses of differentially labeled cross-linking ligands separated in time by 300 s to naïve B cells expressing MHC class II-eGFP (Figure 4.10). This temporal separation between pulses was sufficient to allow the BCRs cross-linked by the first pulse to cluster

on the surface, and to internalize partially. The populations of BCRs that were observed immediately after the second pulse typically were spatially distinct from the aggregated populations of BCRs labeled in the first pulse (Figure 4.11). Correlating the distribution of both populations of BCRs as a function of time after the second pulse (18 cells) indicated that less than half of the two populations of ligated BCRs overlapped immediately after the pulse (45% colocalization) (Figure 4.12). The degree of colocalization increased steadily over time; complete colocalization of the two populations required more than 5 min. These observations suggest that, although some degree of co-staining of BCRs may have occurred with the second pulse, a distinct population of BCRs remained separate from the gross clusters of BCRs formed after the first pulse.

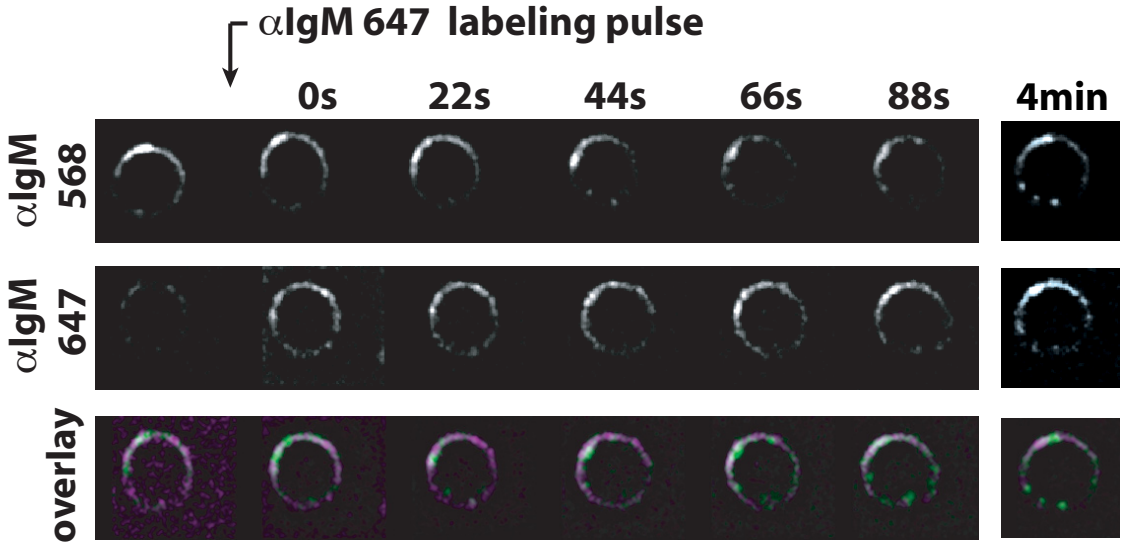

Figure 4.10: Colocalization of temporally-distinct populations of BCRs. Fluorescence micrographs of a B cell labeled with αIgM-Alexa 568 (green, pulse = 30 s) and then after an interval of 300 s, with αIgM-Alexa 647 (red, pulse = 30 s).

The average size of lipid rafts (10–100 nm) makes it difficult to resolve the timing for translocation of cross-linked BCR into rafts by conventional optical microscopy. We chose instead to define a characteristic time (t) representing the first appearance of stabilized patches of BCRs on the surface of the cell larger than 450 nm in size. The median values of t varied as a function of the pulse

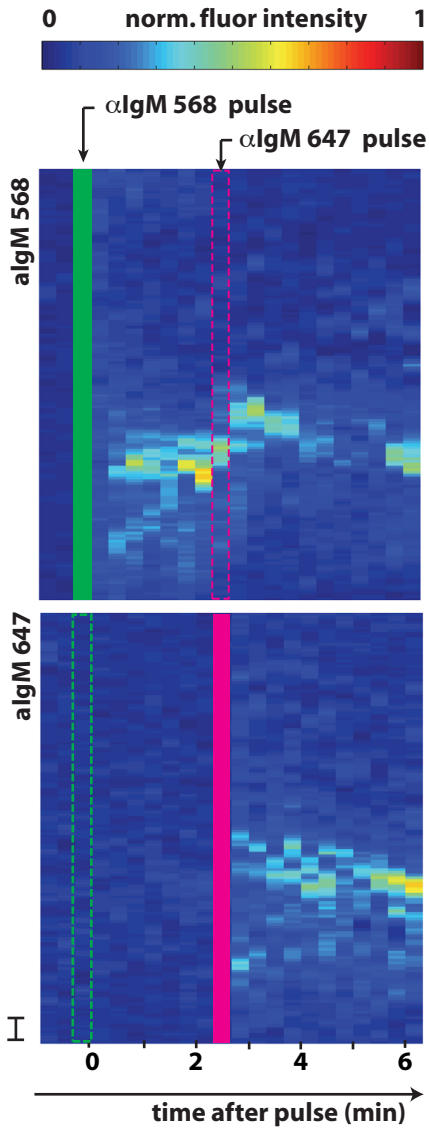

Figure 4.11: Kymographs for the cell shown in (a) for both αIgM-Alexa 568 and αIgM-Alexa 647. The green and pink boxes indicate the position and duration of the applied pulses.

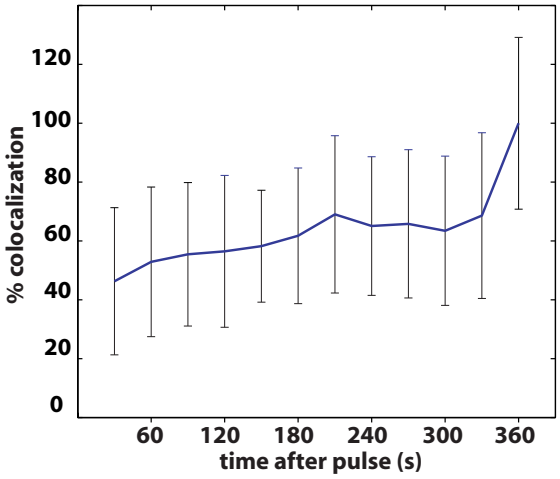

Figure 4.12: Plot of the colocalization of the two populations of labeled BCRs as a function of the time following the second pulse (18 cells).

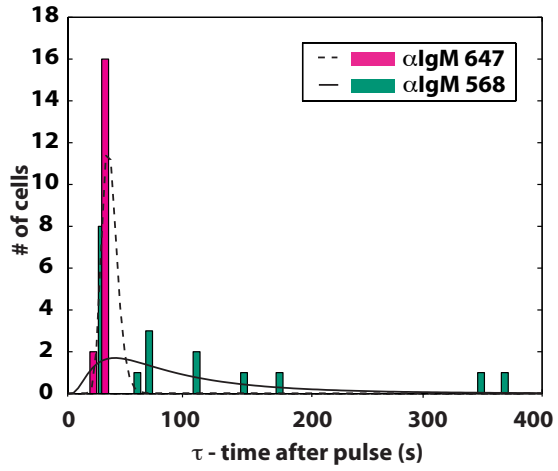

Figure 4.13: Histograms showing the characteristic time (t) at which intense aggregates of BCRs were observed on the surface of the cell following each applied pulse (18 cells).

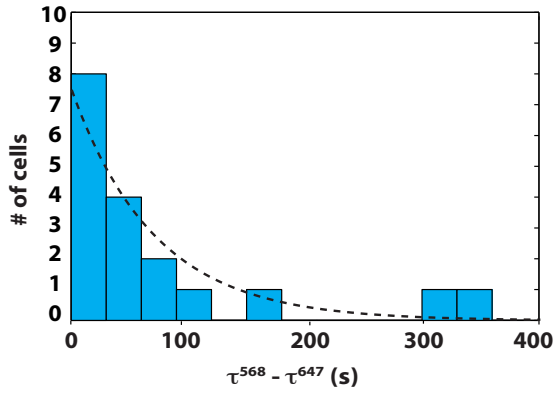

Figure 4.14: Histogram of the difference between the characteristic times for the two populations of labeled BCRs for a given single cell (18 cells).

applied. After the first pulse, labeled BCRs aggregated into microclusters larger than 300 nm in size over a median time of 75.3 s (log-normal fit, geometric std. dev. ±2.3 s, 18 cells) (Fig. 4d). In contrast, the median time for aggregation of BCRs labeled by the second pulse was consistently faster with less variation (median: 35.3 s, geometric std. dev. ±1.2 s, 18 cells). For a given individual cell, the difference between the measured times for aggregation after the first and second pulses showed that the second population of labeled BCRs consistently organized into clusters as fast or faster than the first population (18 cells, [2]).

Class II MHC reorganizes rapidly and independently from BCR

Upon ligation of BCRs, class II MHC molecules at the plasma membrane of naïve B cells also rearrange and internalize into intracellular compartments, which ultimately intersect endocytic compartments containing internalized, ligated BCRs [138] [142]. This process is essential for presentation of the acquired antigen to T cells. To examine the relationship between these two surface-disposed glycoproteins immediately after ligation, we followed the distribution of class II MHC-eGFP on the surface of naïve cells after exposure to a pulse of cross-linking anti-IgM (Fig-

[2] uch-fig4e

ure 4.15). Following ligation, the labeled BCRs and class II MHC appeared to organize on the surface independent of one another: the clusters that formed for one exhibited a low degree of colocalization with the other over time (Figure 4.16). Colocalization of these two populations appeared to increase gradually over time (20–45%), but never approached that observed for the two distinct populations of labeled BCRs. This observation suggests that the two receptors are not closely associated in the plasma membrane of a naïve B cell, and that following ligation and activation, the receptors partition largely into distinct microclusters for internalization.

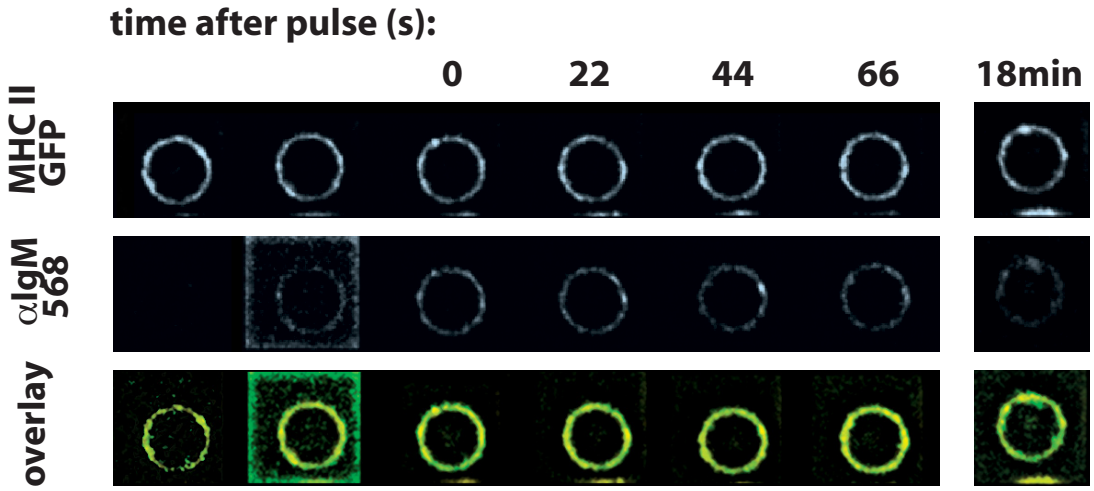

Figure 4.15: Colocalization of class II MHC with labeled BCRs. Fluorescence micrographs of a B cell labeled with αIgM-Alexa 568 (green) and expressing class II MHC-eGFP (yellow). The pulse of αIgM-Alexa 568 was applied for 30 s at the time indicated.

Following a pulse of anti-IgM, surface-disposed MHC class II reorganized into distinguishable aggregates in a characteristic time (tMHC) that was shorter, and less variable, than that observed for the ligated BCRs (Figure 4.17). The median time was 36.7s (lognormal fit, geometric std. dev: 1.2, 18 cells). Examination of the difference between tBCR and tMHC on the same cell further indicated that the time for aggregation of BCRs was consistently slower than that for MHC class II (Figure 4.18, 18 cells). These observations suggest that cross-linking ligation of the BCR (and not subsequent BCR microcluster formation) is sufficient to trigger signaling events that lead to

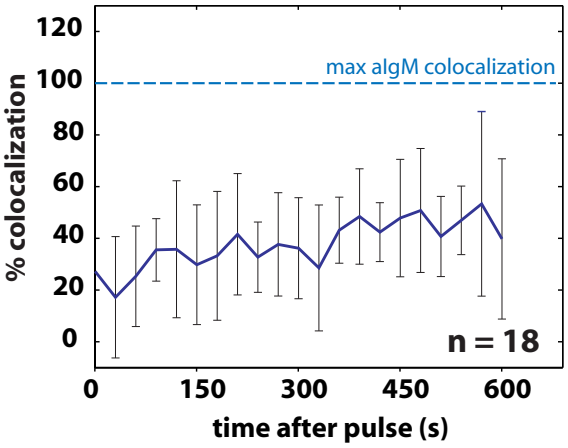

Figure 4.16: Plot of the colocalization of the population of labeled BCRs with the class II MHC-eGFP as a function of the time following the pulse. The average value was calculated from 18 cells.

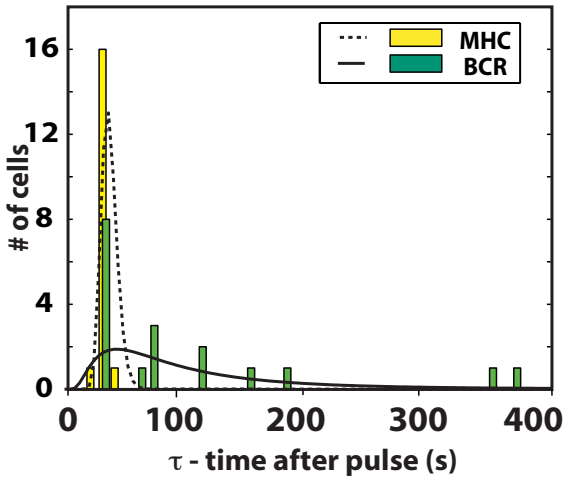

Figure 4.17: Histograms showing the characteristic time (t) at which intense aggregates of BCRs or class II MHC-eGFP were observed on the surface of the cell following the applied pulse (18 cells).

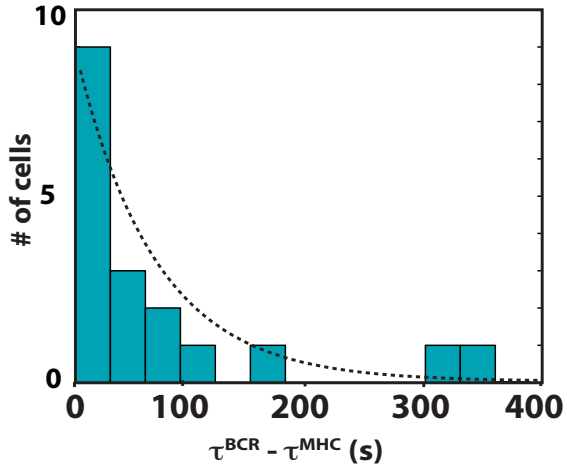

Figure 4.18: Histogram of the difference between the characteristic times for the populations of labeled BCRs and class II MHC-eGFP for a given single cell (18 cells).

the reorganization, and eventually internalization, of both class II MHC and BCRs.

Dynamic rearrangement of BCR and MHCII requires cholesterol

Studies of the BCR by both biochemistry and optical microscopy have indicated that cross-linked BCRs translocate into lipid rafts, which are liquid-ordered microdomains rich in cholesterol [138] [142] [152]. To test whether depletion of cholesterol from the plasma membrane affected the mobility and dynamic rearrangement of BCRs on the surface of naïve B cells, we first exposed cells immobilized in the microchannels to methyl-b-cyclodextrin (MBCD)—a cyclic oligosaccharide that binds cholesterol to deplete it from the plasma membrane and disrupt lipid rafts, and then applied two pulses of cross-linking anti-IgM separated in time by 5 min. Treatment of cells with MBCD immediately disrupted the distribution of MHC class II-eGFP in the plasma membrane, and large clusters were present on the surface of the cells when first observed. The populations of BCRs, labeled by the pulses of anti-IgM, were highly colocalized (70%, 5 cells) with the MHC class II (Figure 4.19 and Figure 4.20).

The population of BCRs labeled by a second pulse also highly colocalized with the first popu-

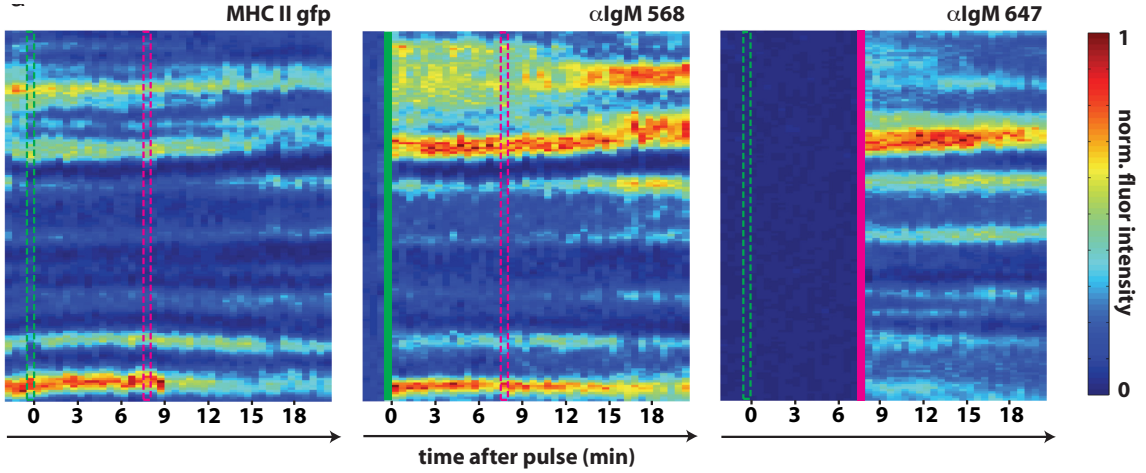

Figure 4.19: MBCD treatment of naïve B cells. Kymographs for a representative cell for MHCII-eGFP, αIgM-Alexa 568 and αIgM-Alexa 647. The cells were treated for 5 min. prior to imaging with 10 mM MBCD at 37°C. The green and pink boxes indicate the position and duration of the applied pulses.

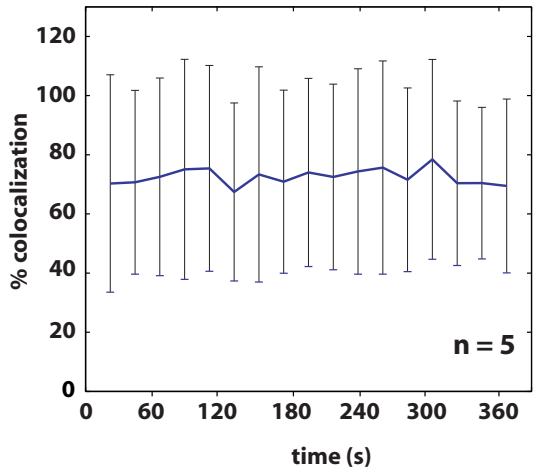

Figure 4.20: Plot of the colocalization of the two populations of labeled BCRs (αIgM-Alexa 568 and αIgM-Alexa 647) as a function of the time following the second pulse. The average value was calculated from 7 cells.

lation. All together, the populations of BCRs and MHCII showed little or no dynamic variation in time with little or no change in the degree of colocalization. These data starkly contrast the distribution and mobility of both BCRs and MHC class II observed on untreated naïve cells, and indicate that, as expected, depletion of cholesterol substantially alters the fluidity of the plasma membrane as well as the distribution of both BCR and MHC class II.

4.3 Discussion

The microfluidic system engineered for these experiments made it possible to examine for the first time the behavior of the BCRs on primary naïve B cells after sequential stimulation with subsaturating quantities of cross-linking ligands or non-specific drug treatments. These experiments provide insight to the early dynamics of the reorganization of both BCRs and MHC class II at the plasma membrane upon limited exposure to cross-linking ligands, and also, on the role of BCR signaling in MHC class II-mediated antigen presentation.

The observed timing of the aggregation of BCRs is consistent with reports on other early events in activation of the signaling cascade. Phosphorylation of Igα and Igβ, a heterodimer component of the BCR complex [153], by the Src-family kinase Lyn in cholesterol-dependent, detergent-insoluble membrane domains is thought to stabilize the partitioning of the BCR into those microdomains [138]. In B cell lines, quantitative FRET imaging has shown that Igα and Lyn closely associate 20–30 s after exposure to a cross-linking ligand [148]. Recruitment of a second kinase, Syk, to the phosphorylated Igα and Igβ happens, on average, 40 s after ligation of the BCR [150]. For naïve B cells exposed to a pulse of ligand, the average time for ligated BCRs to aggregate into intensely fluorescent clusters on the plasma membrane of a cell was 115 s.

Our results suggest that B cells exposed to activating ligand are poised for a more rapid or sustained response upon additional engagement of unligated BCRs than exhibited in the first encounter. The initiation of signaling cascades mediated by the phosphorylation of Igα and Igβ is thought to promote the assembly of rafts containing ligated BCRs into clusters, which also asso-

ciate with the cytoskeleton. The wide distribution of times observed for the gross aggregation of BCRs on individual cells after an initial exposure to ligand (30 s – 380 s) indicates that the process of clustering is a stochastic one, and that the formation of stable aggregates depends on the gradual accumulation of requisite kinases, adaptor proteins, and cytoskeletal elements at the plasma membrane.

In contrast, the narrow distribution of times observed for the rearrangement of ligated BCRs after a second pulse (30–60 s) suggest two possible mechanisms: 1) the components involved in the activation of the signaling cascade may remain at the plasma membrane to allow rapid formation of new signaling complexes upon additional cross-linking of unligated BCRs, or 2) the rate for lateral diffusion of BCRs in the membrane may increase as ligated BCRs are sequestered into cholesterol-rich microdomains. The first suggests a rheostat-like mechanism that would permit incremental increases in signaling that are proportional to the amount of antigen accumulated by a B cell.

Although the scarcity of cytoplasmic space in naïve B cells made it difficult to score internalization of BCR quantitatively, it is worth noting that visual inspection of the data for several cells suggested that the rate of internalization of the population of BCRs labeled by a second pulse of cross-linking ligand was slower than that for the first labeled population. This qualitative observation, in conjunction with the observation by others that phosphorylation of the ligated BCR hinders its internalization [147] [148], further supports the hypothesis that the auxiliary components necessary for stabilizing cross-linked complexes are retained near the surface following their initial recruitment.

The relationship between the signaling cascade activated upon ligation of the BCR and the internalization of aggregated BCR complexes for subsequent antigen presentation via class II MHC has been poorly defined. Biochemical and optical analysis of these two processes have suggested that they are predominantly independent events [147] [148]. B cells expressing class II MHC-eGFP and exposed to a single pulse of cross-linking ligand showed that class II-eGFP reorganized rapidly at the surface after ligation of the BCR, and that the distribution of times for this process was

narrow. The fast response suggests that ligated BCRs do not need to form large, stable microclusters on the surface of the cell to induce the internalization of class II MHC.

In conclusion, the experiments presented here demonstrate the utility of simple microfluidic devices for manipulating the microenvironment around living primary B cells in real time in order to explore the dynamics of their surface-bound receptors. The ability to resolve different populations of identical receptors on the surface by this method has made it possible to distinguish the initial events following ligation of a limited number of BCRs from the events that follow upon additional accumulation of cross-linking ligands. Further exploration of the early dynamics of the activation of B cells with this technology should improve the understanding of the ordering and timing of the initial signaling events, and the relationship between signaling and internalization of BCRs.

4.4 Methods

Soft lithography

Soft lithography comprises a set of techniques for microfabrication based on the use of biocompatible elastomeric materials [154]. Patterns of interest are printed and molded in bas-relief using elastomeric stamps. This approach offers the ability to control the structure of surfaces and to pattern them with proteins and other moieties, to fabricate channel structures appropriate for microfluidics, and to pattern and manipulate cells in three-dimension. For the relatively large feature sizes used in biology ($> 50\mu m$), production of prototype patterns and structures is convenient, inexpensive, and rapid.

Generally, microfluidic systems are fabricated using an elastomeric material, poly (dimethylsiloxane) or PDMS, starting from a patterned silicon wafer employed as a mold. Features of the chip (eg. a network of microfluidic channels) are modeled with computer aided design (CAD). This design is converted into a transparency by a high-resolution printer which, in turn, is used

as a photolitographic mask to create a master in positive relief. PDMS cast against the resulting mold yields a polymeric replica containing all the designed features in bas-relief. The chips are then oxidized in an oxygen plasma, in order to allow the surface of the chip to seal tightly and irreversibly when brought into conformal contact. For instance, oxidized PDMS seals irreversibly to other materials used in microfluidic systems, such as glass or silicon. Oxidation of the PDMS has the additional advantage that it yields channels whose walls are negatively charged when in contact with neutral and basic aqueous solutions, thereby rendering the chip hydrophilic and thus biocompatible, ideally suited for a range of biomedical applications [155].

Fabrication of microchannels

The microchannels were fabricated in poly(dimethylsiloxane) (PDMS, Sylgard 184, Dow Corning) using photolithography and replica molding. One layer of photoresist (SU–8–100, Microchem, Newton, MA) was patterned on a 3" silicon wafer using a transparency photomask (CAD/Art Services, Bandon, OR) to produce a master with a positive relief pattern of the microchannels. To facilitate removal of PDMS in subsequent steps, the masters were silanized by treatment with (tridecafluoro–1,1,2,2-tetrahydrooctyl)–1-trichlorosilane (UCT, Bristol, PA) in a vacuum desiccator for 1 h. PDMS was cast onto the master, cured for 2 h at 60°C, and peeled away. Inlets and outlets were punched into the device using a homemade bore (16G) or a cork borer (5 mm). The microchannels were then treated with an oxygen plasma (PDC–32G, Harrick, Ithaca, NY) for 30 seconds and sealed to a coverglass (24 mm x 50 mm, Corning, Corning,NY); this process also sterilizes the device. The sealed device was filled immediately by capillarity with sterile phosphate buffered saline (PBS, Gibco, Grand Island, NY).

Reagents, Cell Culture, and Mice

Naïve splenic B cells from B6 MHC Class II-GFP+/+ mice or MD4 MHC Class II-GFP+/+ mice were purified by depleting the splenocytes of CD43- or CD11c-positive cells using antibody-conjugated

microbeads (Miltenyi Biotec). Alexa 568-Tfn and Alexa 647-Tfn (Invitrogen) were used for characterization of the pulses administered in the device. Alexa 647-anti-IgM and Alexa 568-anti-IgM (Invitrogen) were employed in pulse-chase experiments with B lymphocytes. Mice employed in these experiments received human care in accordance with the guidelines specified by the Department of Comparative Medicine at MIT under protocol #1005–070–08.

Confocal fluorescence Microscopy

Live cell imaging was performed on a customized inverted microscope (Nikon TE2000E) equipped with a spinning disk confocal head (CSU–10, McBain Instruments) and a 100x TIRF objective. A switchable laser source (Coherent, Santa Clara, CA) was used to excite the fluorophores at 568 and 647 nm excitation wavelengths. Images were acquired with a CCD camera (Hamamatsu, ORCA AG) using Metamorph software (Molecular Devices, Downingtown, PA) for automation of the stage position, z position and time of acquisition. Temperature was maintained at 37°C within a plexiglass chamber surrounding the microscope.

Quantitative analysis of pulses in microchannels

To characterize the concentration profile at the surface of the microchannels during the application of a pulse of soluble ligand, finite element (FEM) simulation results were compared to actual values of fluorescence intensity sampled at an optical plane close to the glass surface. Solution of the partial differential equations (PDEs) by FEM in the 2D and full 3D geometry were obtained in MATLAB (The MathWorks, Natick, MA) and custom code. Models were first created in COMSOL (Comsol, Burlington, MA). Simulations were performed using the MATLAB PDE toolbox. Fluorescence was measured by confocal microscopy in a plane close to the glass surface. Metamorph software was employed to extract the average at every time point. The effect of shear stress on B lymphocytes was quantified employing standard approximation and FEM simulations, over the range of flows (1–4 μL/min) and geometries employed. Applied shear stress peaked at the

time of switching between pulses, reaching values in the order of 1.1 dyne2/cm2. These are values comparable or below measured physiological shear stresses (in the order of 1–10 $dyne^2/cm^2$) and below values found to influence activation in vitro [156] [157].

Pulse-Chase Experiments

Microfluidic devices were coated with poly-l-lysine for 1 h at room temperature and then rinsed in PBS. Naïve B lymphocytes or plasmablasts were transferred into phenol-red free modified DMEM (5%ICS, beta-mercaptoethanol, NEAA, L-Glutamine) and seeded in the device for 20 min at 37°C. Electrostatic attachment to the glass surface did not affect the three dimensional distribution of BCR (data not shown). The microfluidic device was then transferred to the microscope stage, equilibrated at 37°C and connected via plastic tubing (PE–60, BD Biosciences) to syringes containing solutions of ligands and media. To control the flow of media and ligands into the device, we employed syringe pumps (GeniePlus, Kent Scientific), in turn controlled by LABVIEW (National instruments, USA) software equipped with custom routines. Once the flow profile stabilized, multi-dimensional acquisition was started in Metamorph software and synchronized to the program of pulses controlled by LABVIEW.

Image Processing and Analysis

Raw fluorescent data collected with Metamorph (Molecular Devices, USA) was imported into MATLAB and organized into time series. A single exponential was fit to the photobleaching curve previously determined (data not shown) and used to determine the correct multiplicative factor to adjust total fluorescence at every time point. An annular mask encompassing the entirety of each cell membrane was created individuating circular structures by a Hough transform-based algorithm. Fluorescent intensity across the annulus was sampled as a function of angle, averaging over the finite thickness of the annulus itself. Kymographs were generated by normalizing the fluorescence to the maximum value recorded in each experiment and arranging fluorescent data as a

function of angle and time. The same time series was employed to determine the first occurrence of significant levels of clustering, defined as the first time point where fluorescent peaks of width $>$ 450nm (linear circumference) are observable and accompanied by monotonic increases in overall signal strength (in order to discriminate significant labeling from simple variations in noise levels). The time was recorded for each cell and in each channel. Lag was calculated by subtracting the characteristic time in each channel for each cell. Time data was fitted by a lognormal distribution (two parameters, μ and σ), for which geometric mean (exp[μ]) and geometric standard deviation (exp[σ]) were reported. Colocalization was measured by Pearson's correlation coefficients, employing normalized raw fluorescent data. Similar analysis was applied to a 3D reconstruction to confirm the lack of apparent bias in the three-dimensional distribution of ligand binding at the surface. Full z stacks of a lymphocyte subjected to a pulse of ligand concentration were collected over time. A set of differentially oriented two-dimensional slices were obtained by reslicing the reconstructed volume (Volocity, Improvision). There was minimal variance in the distribution of intensity amongst all slices compared (data not shown).

Equations describing the microchannel flow

The flow in the microfluidic device can be fully described by a system of coupled partial differential equations:

$$\frac{\delta L(x,z,t)}{\delta t} + \vec{v}\nabla L(x,z,t) = D\nabla L(x,z,t)$$

$$\nabla \cdot \vec{v} = 0$$

$$\rho\frac{D\vec{v}}{Dt} = \rho g - \nabla P + \mu\nabla^2\vec{v}$$

where L is concentration of ligand, v is the velocity profile in the channel, D is the diffusive constant, ρ is viscosity. The coordinate x-axis is oriented along the length of the channel, the y-axis is oriented across the width and the z direction describes the height from the bottom glass surface

(where cells are plated).

Important nondimensional numbers that describe the flow include Re and Pe:

$$Re = \frac{\rho v_{avg} h}{\rho}$$

$$Pe = \frac{v_{avg} h}{D}$$

For flow in microchannels Re is order 1 or less and flow is laminar. We designed our system to have high Pe number, in the order of 105, the transport process is dominated by convenction with diffusion effects playing a role exclusively in a thin region near the boundary.

We limit analysis to the two dimensional (2D) case, solving exclusively the x-z components of the above PDEs. For two dimensional flow, the velocity profile is given by

$$\vec{v}(z) = \frac{1}{2\mu}\left(-\frac{\delta p}{\delta x}\right) z(z-h)$$

where h is the height of the channel and p is pressure. Shear stress for flow in a 2D rectangular cross section channel is then given by

$$\tau_x = \mu \frac{v}{z}|_{z=0} = \frac{12\mu\vec{v}}{h}$$

Boundary conditions

We employed the following set of standard boundary conditions

$$L(x, z, 0) = 0$$

$$L(0, 0, t) = L_0 \left(u(t - t_0) - u(t - t_1)\right)$$

$$D\frac{dL}{dz}|_{z=0,h} = 0$$

where u(t) is the step function, Lo is the concentration in the segregated stream of ligand and

to, t1 are the beginning and end time of the pulse respectively. Here we assumed that the pulse has uniform concentration at the inlet, an assumption that was corroborated by full 3D simulation of the concentration profile.

Following our design, for the range of Peclet number (Pe) employed in our device ($\sim 10\hat{}5$), the solution does not depend significantly on the axial position in the channel or the imposed initial concentration.

Reaction at the cell-populated surface

The boundary conditions at the bottom of the channel merits further discussion. We decided to solve numerically the given set of PDE maintaining zero flux at the surface, although in reality a reaction is occurring at this surface, due to the presence of cells. The boundary condition is given by

$$D\frac{dL}{dz}|_{z=0,h} = k_f R_T L(x,z,t)$$

from which we can observe how reactions of different magnitudes at the surface, with different reaction rates k, will affect the concentration distribution of the ligand in the channel quite dramatically.

In fact, for fast enough reaction rates, the effective concentration at the surface approaches zero. The extent of this effect can be gauged by calculating the Dahmkohler number for the system of interest, which gives the relative contribution of the mass transfer process to the surface, compared to the reaction occurring at the surface

$$Da = \frac{diffusion}{reaction} = \frac{h^2 k_f}{D R_T}$$

Yet, as long as the effective ligand concentration remains at levels below receptor saturation, our solution remains a good estimate of the effective ligand concentration. That is, to a first approximation, the value in which we are interested - the magnitude of the overall reaction term - does not

change significantly when calculated directly from the boundary condition or when extracted from the concentration profile at the surface in the zero-flux case.

Importantly, by employing the zero-flux approach, we obtain a concentration profilewhich we can apply in the analysis of a variety of different kinetic reactions and hence different cell systems, thereby increasing the applicability of the numerical solution.

Cells are exposed to pulses at sub-saturating concentrations of ligand

To better understand the effects of a transient concentration pulse on cell labeling, it is necessary to quantify the extent of the labeling reaction occurring on the cell surface. In particular, it's necessary to determine the operating regimes for which the labeling concentration does not saturate cell surface receptors.

The absorbed concentration depends on the concentration of the ligand, the binding constant, and the type of reaction. A reasonable estimate of the receptor occupancy can be obtained assuming a reversible reaction at the channel surface, where ligand binds receptors to form complexes at the cell surface.

In our system the reaction normally under consideration is between the bivalent ligand anti-IgM and BCR receptors on the surface of B lymphocytes. However, rather than solving the highly nonlinear problem of bivalent ligand, we employed a much simpler monovalent ligand description to obtain a first indication about the operating regime of the device. In the case under consideration the approximation is a particularly good one, due to the high binding constant of these two ligands; we can safely exclude serial engagement and assume in our analysis that ligand is sequestered rapidly from the freely diffusing fraction upon crosslinking and internalization.

The binding process in the presence of ligand is highly transient ($\sim$10–20 seconds) and classic equilibrium solutions do not apply (for the given ligands equilibrium is only reached for time scales in the order of minutes). We can obtain an estimate of maximum receptor occupancy, by evaluating the transient solution at the point in which it reaches its maximum value, which correspond in this case to the end of the pulse. For a simple monovalent reversible reaction $R + LC$, with initial

receptor occupancy equal to zero, the equation is given by:

$$\frac{C(t)}{R_T} = \frac{L_0/K_D}{1 + L_0/K_D}\left(1 - exp\left(-\left(1 + \frac{L_0}{K_D}\right)k_r t\right)\right)$$

where L_0 is the ligand concentration associated with the pulse (which we assume constant), K_D is the dissociation constant, R_T is the total number of receptors and k_r is the reverse binding constant.

We decided to treat the concentration of ligand as constant for the time course of the pulse, given the rapid transient, the relatively uniform concentration profile over the length of the pulse and the negligible depletion effects. We also assume ligand depletion effects are negligible given the small number of total receptors, when compared to the number of ligand molecules in a pulse.

$$\left(\frac{n}{N_a v}[Complex]\right) \ll [L_0]$$

where n is the cell density and L_0 is the imposed ligand concentration. We also select to ignore diffusional effects at the receptor surface, in order to obtain a more conservative estimate. Maximum receptor occupancy is therefore given by:

$$R_{occupancy} = \frac{C(t = \delta t)}{R_T}$$

where Δt is the effective duration of the imposed pulse.

It is important to notice how for the time scales of interest (below 1 min), few combination of parameters will give a receptor occupancy larger than 50%. For the antiIgM –BCR interaction, where the polyclonal antibody is a nanomolar binder, we have $k_f = 10^4 \frac{M}{s}$, $k_r = 10^-5 \frac{1}{/} s$, $K_D = 10^-9$ M and $R_T = 10^5$. For the range of concentrations we employed, L_0 = 60–120 nM, the maximum receptor occupancy is in the range 1–11%, well below the saturation regime.

Chapter 5

Profiling antibody responses by multiparametric analysis of primary B cells

Craig M. Story*, Eliseo Papa*, Chih-Chi Andrew Hu, Jehnna L. Ronan, Hidde L. Ploegh, and J. Christopher Love.

* These authors contributed equally

This chapter is presented as it originally appeared in PNAS 2008 105 (46)[1]

Corresponding supplementary material is appended

[1]doi:10.1073/pnas.0805470105

Abstract

Determining the efficacy of a vaccine generally relies on measuring neutralizing antibodies in sera. This measure cannot elucidate the mechanisms responsible for the development of immunological memory at the cellular level. Quantitative profiles that detail the cellular origin, extent and diversity of the humoral (antibody-based) immune response would improve the assessment of vaccines, and thus the development of new vaccines. Here we describe a novel approach to collect multiparametric datasets that describe the specificity, isotype, and apparent affinity of the antibodies secreted from large numbers of individual primary B cells ($10^3 - 10^4$). The antibody/antigen binding curves obtained by this approach can be used to classify closely related populations of cells using algorithms for data clustering, and the relationships among populations can be visualized graphically using affinity heatmaps (AffiMaps). The technique described was employed to evaluate the diversity and cellular origins of the antibodies generated during an in vivo humoral response to a series of immunizations designed to mimic a multipart vaccination. Profiles correlating primary antibody-producing cells with the characteristics and diversity of their secreted antibodies should facilitate both the evaluation of candidate vaccines, and broadly, studies on the repertoire of antibodies generated in response to infectious or autoimmune diseases.

5.1 Introduction

Antibodies generated by a natural immune response or by a vaccine are important for limiting the ability of a pathogen to infect the host [158] [159]. In many instances, the concentration of those antibodies circulating in the blood provides a surrogate measure for the efficacy of a vaccine. In vitro immunoassays, such as the enzyme-linked immunosorbant assay (ELISA) or opsonophagocytic assay (OPA), can determine the ability of antibodies from sera to neutralize the appropriate virus or toxin, to induce anti-microbial activity, or to simply bind a specific component in the vaccine. These measures provide some insight into the degree of immunity afforded, but obscure the relative contributions from individual antibodies. Furthermore, analysis of the sera does not reveal anything about the B cells and plasma cells responsible for producing the antibodies. Such detailed knowledge of the molecular characteristics and cellular origins of unique antibodies raised during a humoral response would benefit the design and evaluation of new vaccines [Wack2005][]. The attributes of an antibody, including its isotype and affinity for an antigen, can distinguish it from others with similar specificities. It is challenging, if not impossible, however, to separate individual antibodies present in polyclonal sera in sufficient quantity to produce a complete assessment of their molecular characteristics. Instead, a number of techniques have been developed to examine the antigen-specific B cells that produce these antibodies. Methods for measuring the frequency of antigen-reactive B cells include the Jerne plaque assay [160], the splenic focus assay [161], a variant of the plaque assay that uses colonies of cells immobilized on discs of filter paper [162] [163], the enzyme-linked immunospot (ELISpot) assay [164], and fluorescence-activated cell sorting (FACS) [165] [166]. None of these methods, however, can by itself define a comprehensive set of correlated characteristics suitable to describe the diversity of antibody-producing B cells responding to a vaccine or other immunological challenge.

One important characteristic of an antibody that remains difficult to assess in a rapid and high-throughput manner is its apparent affinity for a given antigen. Presently, assessing the affinity of a monoclonal antibody involves the preparation of antibody-secreting cell lines, the production and

purification of sufficient quantities of the antibody, and finally characterization by ELISA or surface plasmon resonance (SPR). The transformation of primary cells into antibody-secreting clonal lines requires either immortalization of cells by fusion or viral transduction, or molecular cloning. Both processes are relatively inefficient at capturing the entire repertoire of clones present from the primary cells, and as a result, the frequencies measured may vary from the distribution found ex vivo. These processes also are time-consuming, and thus further impose a practical limit on the total number of clones evaluated. Here, we report a method to estimate an apparent affinity/avidity for the antibodies produced by individual primary cells that does not require immortalization of the cells. The approach uses a soft lithographic technique, called microengraving, to produce a set of replicate microarrays of antibodies from a population of cells [167]. The precision of the binding curves measured for the antibodies produced by each cell is sufficient to organize related cells into distinct groups using common statistical algorithms for data clustering. We also demonstrate the first example of datasets that simultaneously describe the antigenic specificity, isotype, and apparent affinity of the antibodies produced by individual primary B cells ($\sim 10^4$) generated in mice given a series of immunizations designed to mimic a multipart vaccination. Together, these data constitute a detailed profile of the individual B cells contributing to the humoral immune response with a resolution not previously possible.

5.2 Results

Microengraving is a soft lithographic technique for printing protein microarrays where each spot in the array comprises the proteins secreted by a single cell. The method uses an array of microscopic wells to segregate individual cells spatially into subnanoliter volumes. A typical array comprises ~80,000 microwells arranged over an area of $\sim 20 \times 60\ mm^2$ on a polymeric slab; each well is 0.1 nL in volume ($50 \times 50 \times 50\ m^3$). Cells settle by gravity from a suspension into the microwells and weakly adhere to the bottoms of the wells at a density of ~1 cell per well. The array is then placed in contact with a glass slide, and the antibodies secreted by each cell are captured on the surface

of the slide which is uniformly coated with a secondary antibody. The arrangement of captured antibodies on the resulting microarray matches the arrangement of cells in the microwells. This printing process can be repeated with the same set of cells to produce multiple replicate antibody microarrays (see Figure 5.1 Figure 5.2 for an illustration) [167].

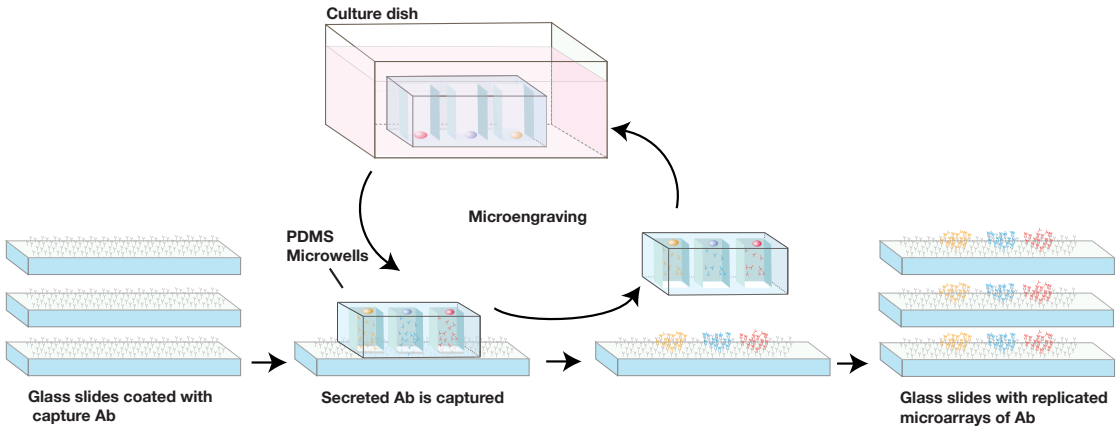

Figure 5.1: Experimental method for assessing the affinities of antibodies produced by single cells. Schematic illustration of the microengraving process for generating multiple replicates of microarrays of antibodies from one population of cells. Cells are loaded into the microwells molded into the PDMS slab, and then used to print onto glass slides functionalized with capture antibodies. After producing microarrays, the microwells containing cells are transferred to media for subsequent enumeration by immunofluorescence.

We reasoned that a set of 6–7 replicated arrays would be sufficient to determine the specificity, isotype, and apparent affinity of the captured antibodies produced by a given cell in a microwell. One replicate provides data on the isotype of each antibody when interrogated with a collection of differentially-labeled secondary antibodies specific for an isotype, or a subclass (e.g., goat-anti-mouse IgM, IgG1 or IgG2a). The remaining replicates can be used to assess the apparent affinity of each antibody captured on the microarrays for a particular antigen. A binding curve for the antibodies produced by each cell can be constructed by first applying a series of concentrations of antigen (e.g., 10 pM – 100 nM) to a set of replicate microarrays, and then measuring the fluorescent intensities of captured antigen as a function of concentration [168]. Variations in the rates of secretion among cells or in the efficiency of capture among glass slides can both affect the amount

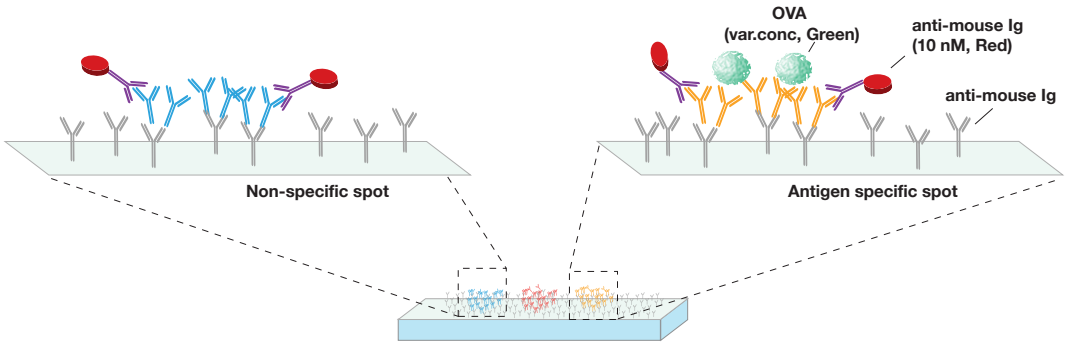

Figure 5.2: Schematic illustration of the labelling scheme for assessing concentrations of captured monoclonal antibody and antigen at a given element in the microarray. Each replicate print is exposed to antigen at one concentration within a specified range (e.g., 10 pM to 100 nM) and a fixed concentration of a labelled secondary antibody (10 nM). Spots containing antibody non specific for the antigen appear in one color only.

of antibody captured from each cell on each replicate. To account for the intra- and interarray variations, we also labeled the arrays with a fixed concentration of an anti-Ig antibody to assess the amount of antibody captured at each element. The ratio of the fluorescence intensities measured for bound antigen to that for the amount of antibody captured (anti-Ig) normalizes the antigen captured at a given concentration and location on the array by the amount of antibody available at that site. Varying the concentration of antigen applied across replicate microarrays, while holding the concentration of the secondary antibody constant for each, yielded a series of ratios used to construct the binding curves for that antibody. An apparent dissociation constant (KD_app) for an antibody from a given cell can then be calculated by fitting these curves.

Single-cell measurements to assess the affinity of monoclonal antibodies.

We used a monoclonal line of hybridomas producing antibodies specific for chicken ovalbumin to test the feasibility of the approach described above for estimating the affinities of antibodies captured from single cells. We deposited hybridomas into an array of microwells, and then produced five replicate microarrays of antibodies from that array of cells. Between each print, the media in

the wells was exchanged by gently rinsing the surface of the microwells. Each microarray of antibodies was then exposed to a different concentration of fluorescently-labelled ovalbumin (10 pM to 100 nM) (see Figure 5.3).

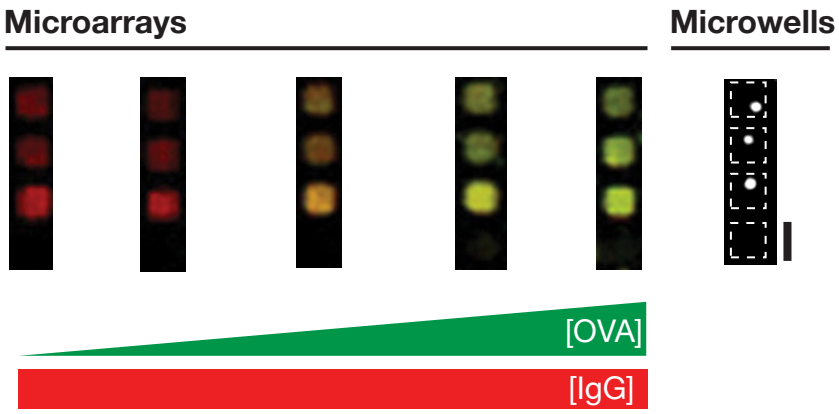

Figure 5.3: Measurements of the affinities of antibodies produced by single cells. Representative composite fluorescent images from five antibody microarrays produced by microengraving labelled with fluorescent antigen, and corresponding region of wells containing anti-ovalbumin hybridomas stained with carboxyfluorescein succinimidyl ester (CFSE). The dashed white boxes indicate the physical edges of each microwell. Scale bar is 50 μm.

The data from each cell was fitted with a langmuir binding curve to estimate KD_app: The median value for 3,461 individual cells was 0.79 nM and exhibited a very narrow distribution (95% bootstrap confidence interval: 0.78, 0.80 nM) ([2]).

Although the precision of the measurement is most critical for distinguishing related cells, we also evaluated the accuracy of the measured value relative to ELISA and SPR. The apparent dissociation constant determined by ELISA using the purified antibody and plate-bound antigen was 0.5 nM (Figure 5.4 insert), and the affinity of the antibody determined by SPR using soluble antigen and bound antibody was 0.11 nM. These data suggest that the measurements are reproducible for a given clone and that the estimated affinity is a close approximation to that determined from bulk quantities of antibody.

[2] pnas-fig1b

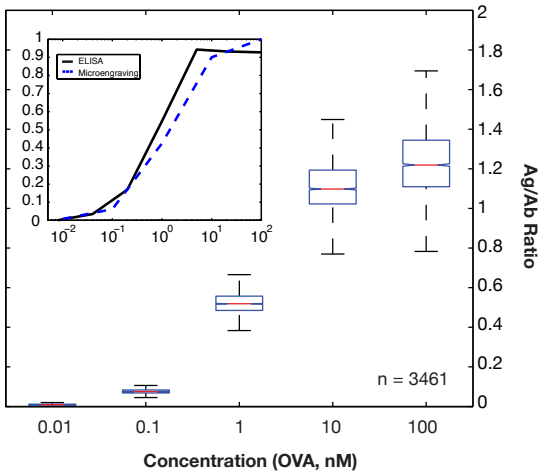

Figure 5.4: Plot of the fractional occupancies measured for 3,461 cells as a function of the concentration of ligand applied to five replicate microarrays. (Inset) The median binding curve and the curve measured by ELISA for the purified antibody are shown.

Binding curves are sufficient to distinguish closely-related hybridomas

The high degree of precision observed for the individual binding curves measured for the anti-ovalbumin hybridomas suggested that it should be possible to distinguish similar clones within a mixed population by comparing the binding curves measured for their antibodies. We repeated the experiment described above using a mixture of three different monoclonal hybridomas producing antibodies specific for mouse major histocompatibility complex (MHC) class I (anti-H–2Kb). Specifically, three hybridoma cell lines were each labelled with unique cytosolic dyes, and then a mixture of these three was deposited onto an array of microwells at a density of $\sim$1 cell per well. We generated seven replicate microarrays by microengraving using this set of cells, and stained them with tetrameric complexes of class I MHC; tetrameric antigen was used to improve detection without obscuring the epitopes recognized by the antibodies, and to increase the strength of the interaction for the low affinity clone (c127). After the final print, all of the cells were stained in situ for DNA, imaged by immunofluorescence, and subsequently matched to the corresponding antibodies on the microarrays (Figure 5.5 and Figure 5.6).

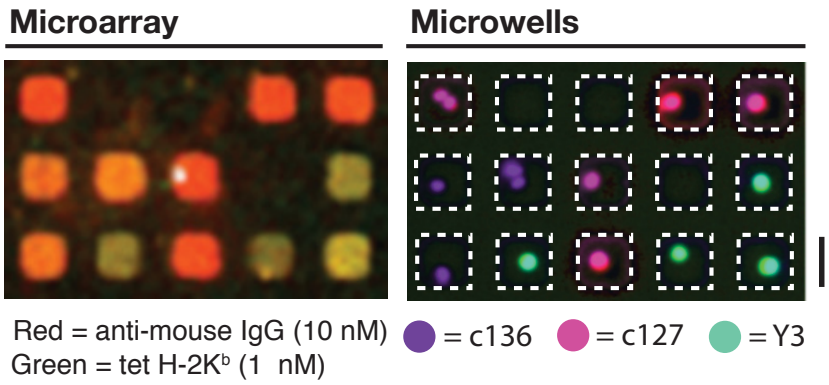

Figure 5.5: Binding curves distinguish similar hybridomas. Representative composite fluorescent images of one microarray and the corresponding cells in microwells (Y3, aqua; c127, magenta; c136, purple). The microarray was imaged after staining with anti-mouse IgG (red) and tetrameric MHC class I (H–2Kb). The cells were labelled with three different cytosolic dyes before loading the microwells, and stained for DNA content after printing. The dashed white boxes indicate the boundaries of each microwell. The scale bar is 50 μm.

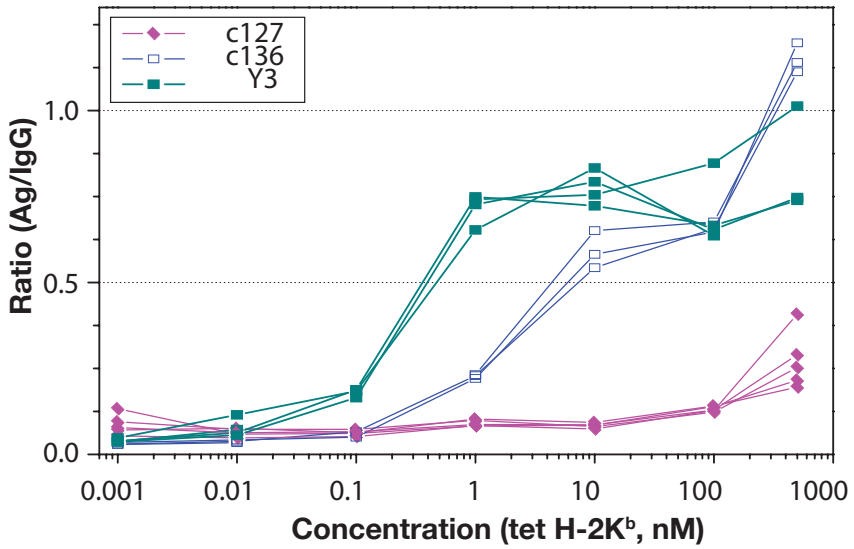

Figure 5.6: A plot showing the fractional occupancy of each element in the region of the microarray shown in Figure 5.5 as a function of the concentration of tetrameric H–2Kb. The colors of the traces correspond with the identity of the clone determined by immunofluorescence.

The experiment yielded a collection of binding curves matched to more than 3,700 single cells. These data allowed us to confirm that the measured binding curves were consistent and distinct for each clone. Apparent avidities were calculated and compared to the anti-ovalbumin hybridoma, and the precision of these estimates was determined by calculating bootstrap confidence intervals (Figure 5.7 and Figure 5.8).

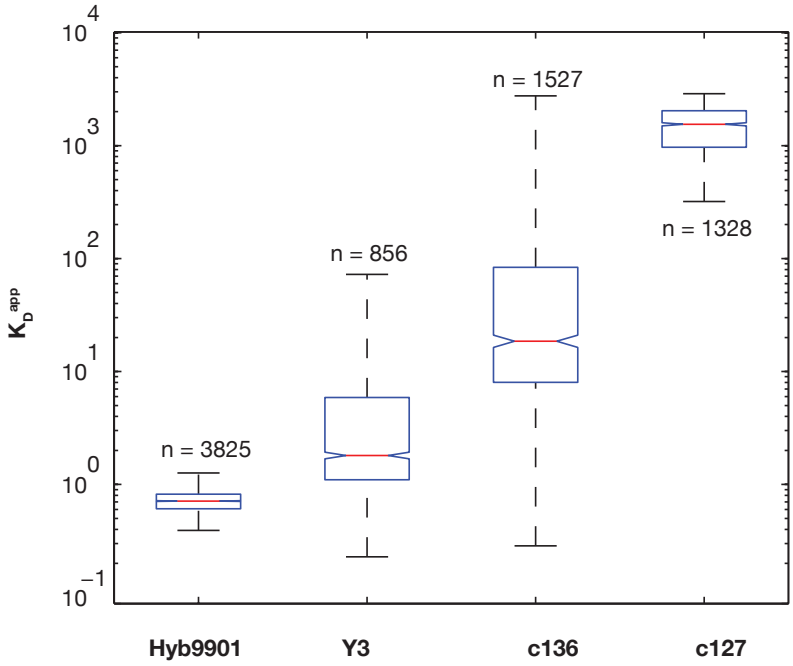

Figure 5.7: Apparent dissociation constants measured in all hybridoma clones employed. (a) Boxplot of apparent dissociation constants obtained by langmuir fit on the microengraving data. Notches show the 95% confidence interval.

The magnitude of the intervals varied for each clone, but in each case, were approximately one to two orders of magnitude less than the calculated value, indicating the high precision of the measurements. To validate that the affinities of the three clones were indeed different, the affinities of each antibody for monomeric class I MHC were determined by SPR (data not shown). The measurements confirmed that the three clones produce distinct antibodies, but the values did differ from those estimated by microengraving. This result was expected since the configuration of

Clone	Median K_D^{app}	95% Confidence Interval
Hyb9901	0.79	[0.78, 0.80]
Y3	1.97	[1.88, 2.11]
c136	19.7	[17.1, 22.3]
c127	1555	[1506, 1615]

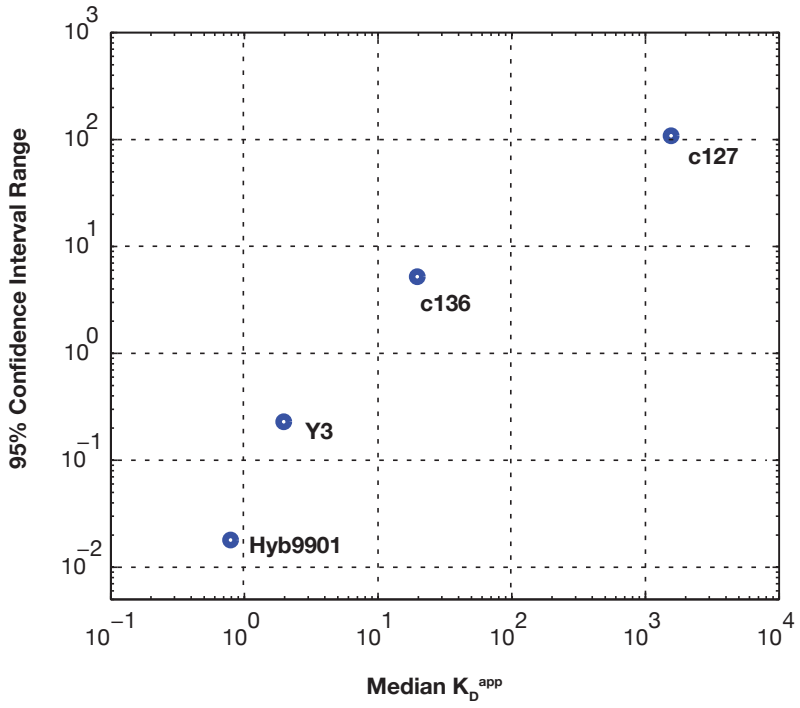

Figure 5.8: Table of the actual numerical values of the median apparent dissociation constant and 95% confidence intervals. (c) Accuracy of the microengraving method over a range of dissociation constants. Confidence intervals (95%) are plotted as a function of measured apparent dissociation constant.

the assay, the differences in valency, and capture antibodies could not be matched across the two techniques. Moreover, our technique – differently than SPR measurements – is affected by the contribution of the non-functional fraction of the secreted antibodies. In order to visualize the entire dataset, it was necessary to develop a compact graphical representation corresponding to the binding curve for each cell (Figure 5.9 and Figure 5.10).

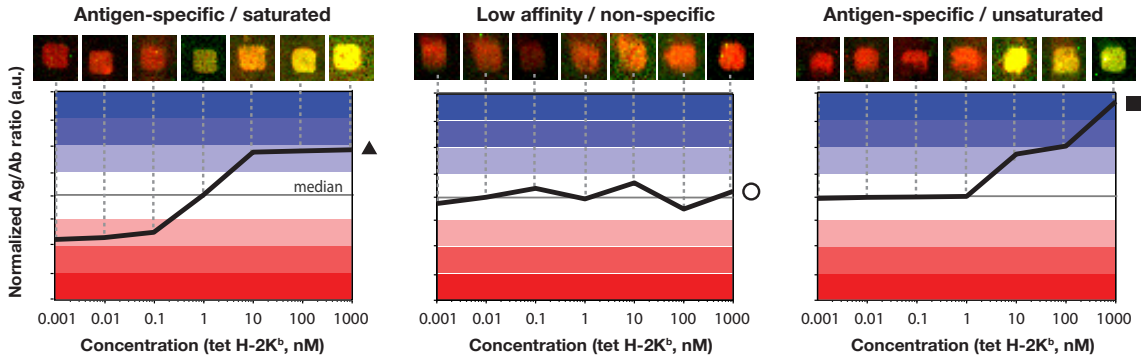

Figure 5.9: Method for constructing an affinity heatmap (AffiMap). Three examples of measured binding curves from the anti-H–2Kb clones are shown: i) antigen-specific antibodies that exhibit saturation of binding (triangle), ii) non-specific/low-affinity antibodies (circle), and iii) antigen-specific antibodies that do not exhibit saturation of binding (square). The composite fluorescent images correspond to each data point in the median-centered curves. The colored bands shown on the plots indicate the colors used in the corresponding density plots in Figure 5.10.

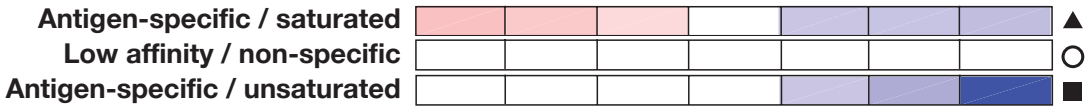

Figure 5.10: Construction of an AffiMap for the three clones in Figure 5.9. Each row represents the binding curve measured for the antibody secreted by one cell as a density plot. Antigen-specific antibodies that do not exhibit saturation show only blue coloring. Non-specific or low affinity antibodies have no coloring.

We centered each curve such that its median value was positioned at zero. The data were then converted to a density plot where points above and below the median were represented in contrasting colors. In this form, the density plot for an antibody with a high affinity for the target antigen

shows a progression of color (from red to blue) with an inversion at the median (white), while the density plot for an antibody with a low affinity is almost completely white. This representation is also convenient because stacking such lines vertically yields a concise picture of the entire dataset. The resulting image can be examined quickly to identify antibodies with measurable affinities (colored lines) and to highlight the relationships between cells with similar apparent affinities for the target antigen. We refer to these visual representations of the affinity measurements collected for a population of cells as 'affinity heatmaps', or AffiMaps.

The differences visually evident from the binding curves measured for the three clones suggested that statistical algorithms for data clustering should be able to classify the cells into related populations. We tested the utility of three unsupervised clustering algorithms for sorting cells into groups of identical clones: k-means clustering, one-dimensional hierarchical clustering, and principal-components analysis (PCA) (Figure 5.11, Figure 5.12 and Figure 5.13).

The data collected by cellular staining provided a reference to assess the accuracy of the assignments made by the computational algorithms. For both k-means and hierarchical clustering, the average sensitivity (ratio of true positives to combined true positives and false positives) was greater than 87% and the average specificity (ratio of true negatives to combined true negatives and false positives) was greater than 94% (see Table 5.1).

Table 5.1: Sensitivity and specificity of clustering algorithms to model data

-	Clone			Measure			
Classifier	Y3	c136	c127	Avg. sens	Avg. spec.	Avg. PPV	Avg. NPV
Actual(stains)	1527	856	1328	-	-	-	-
K-means	1482	831	1398	87.0%	94.3%	87.0%	94.4%
Hierarchical	1617	744	1350	88.3%	95.2%	89.8%	95.6%

Multiparametric profiles of humoral immune responses in mice. The ability to distinguish accurately amongst three monoclonal immunoglobulins specific for the same antigen suggested that it should be possible to examine the diversity in a more heterogeneous population of B cells, such as those involved in a developing humoral immune response. Mice were immunized with chicken

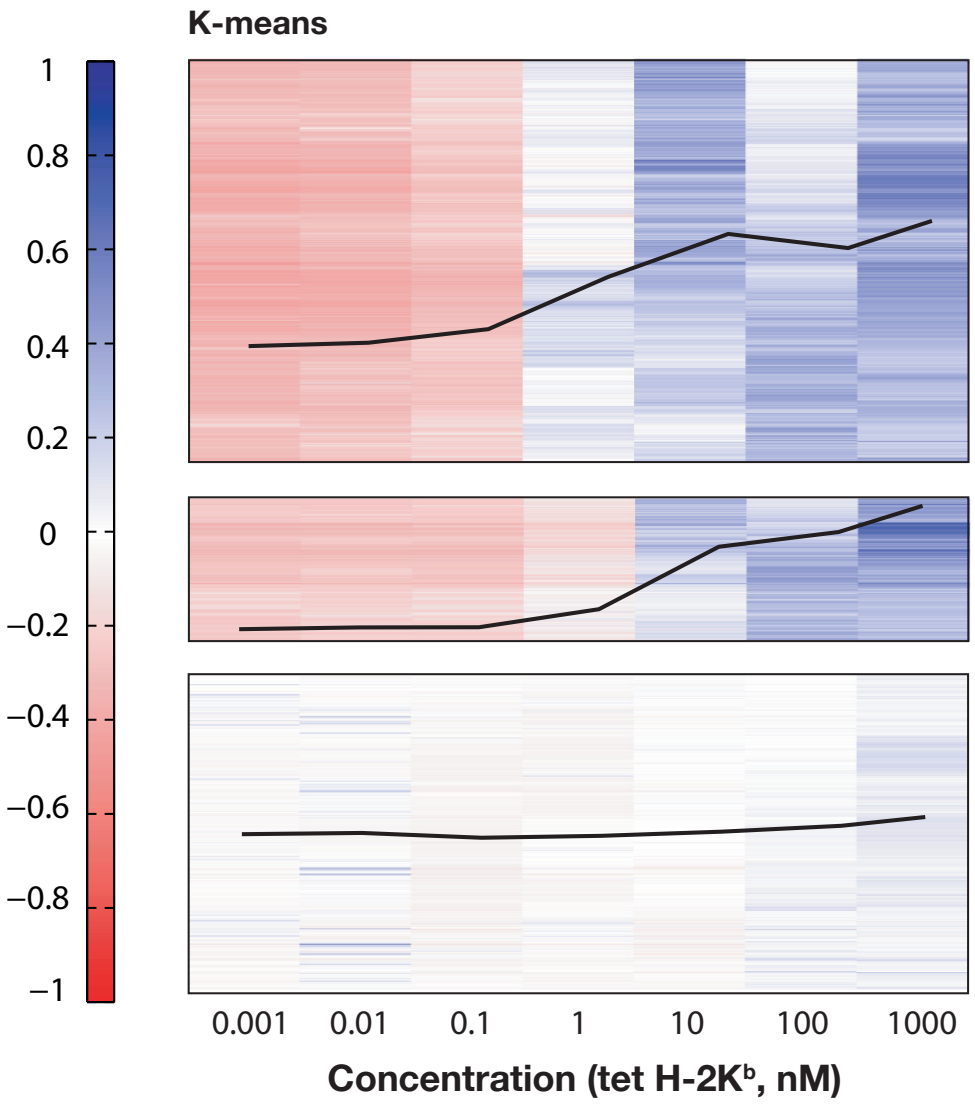

Figure 5.11: AffiMap of anti-H–2Kb hybridomas generated by k-means (n=3) clustering (3,711 cells). The median profiles for each cluster are superimposed (black).

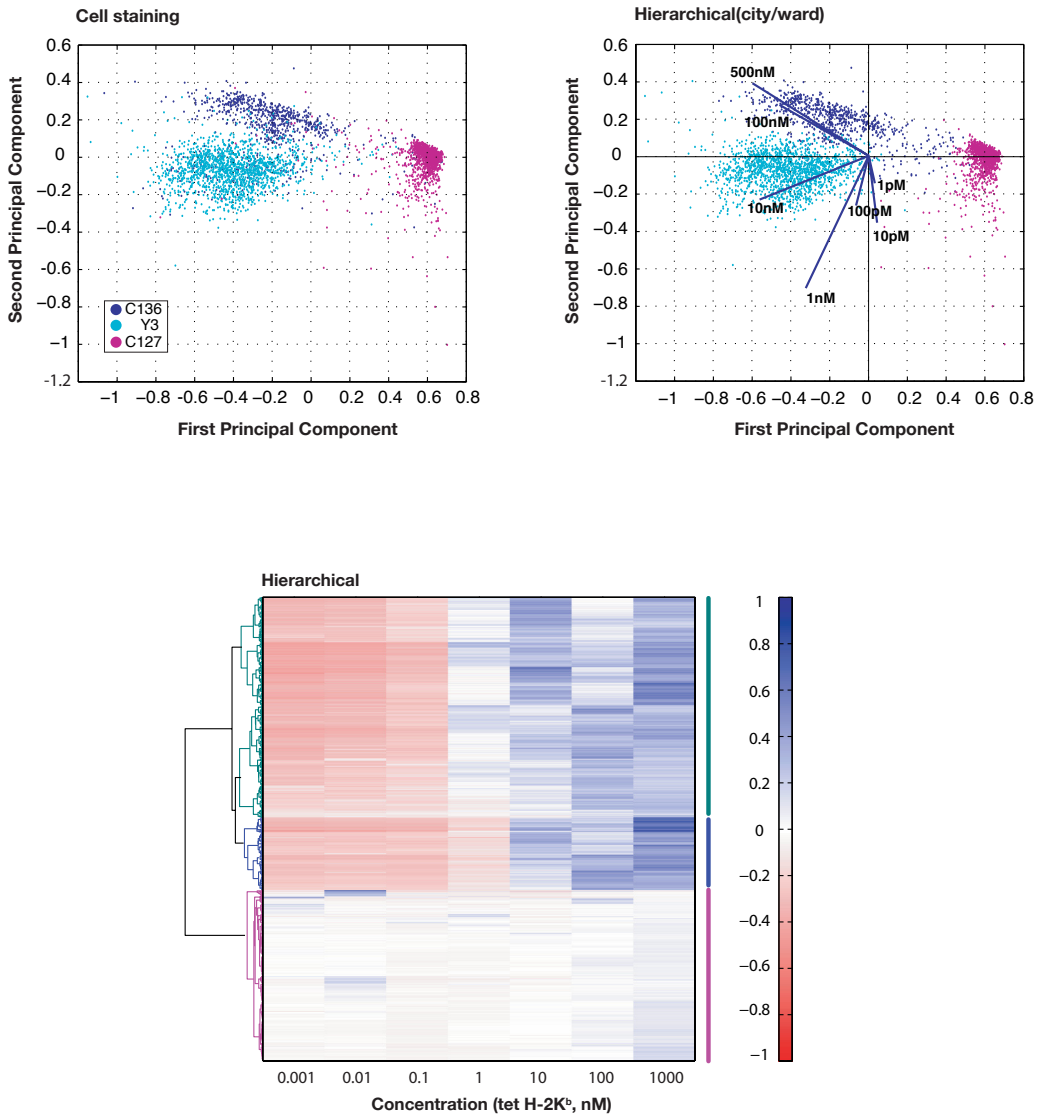

Figure 5.12: AffMap for three anti-H–2Kb hybridomas organized by hierarchical clustering (city/ward linkage). Three dominant subclusters are indicated with colors corresponding to the cytosolic labels in Figure 5.10

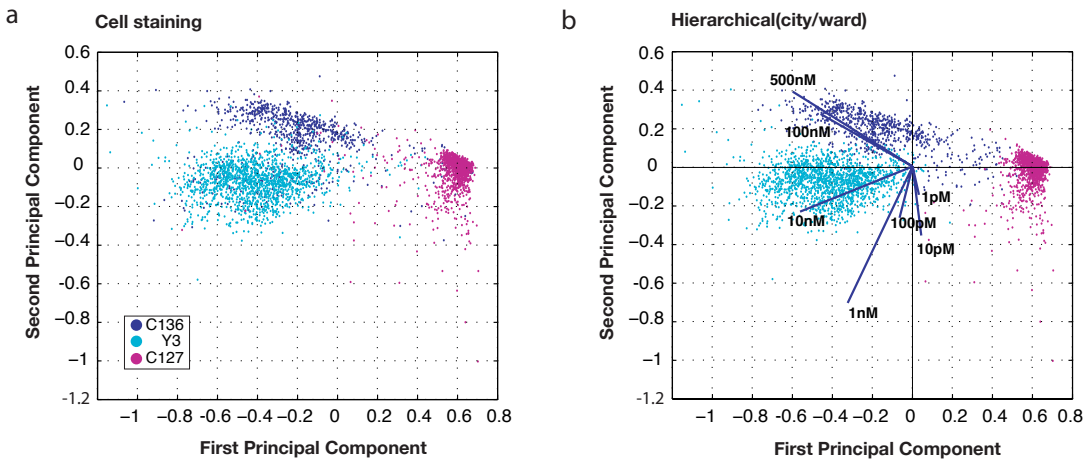

Figure 5.13: Principal-components analysis of affinities measured for three anti-H–2Kb hybridomas. Single cells are plotted as a function of the first two principal components. Colors indicate the identity of the cells assigned (a) by intracellular staining and (b) by hierarchical clustering. The vectors show the two-dimensional projection of the original multidimensional basis of the principal-component space.

ovalbumin and boosted either once or twice (see Figure 5.14).

Splenocytes isolated from these mice were stimulated ex vivo with a mitogen (lipopolysaccharide, LPS) for three days to convert B cells into immunoglobulin-secreting cells. We then characterized these cells by a combination of microengraving and immunofluorescence. The complete dataset collected for each cell comprised 1) DNA staining to assess the number of cells per well, 2) surface-expressed markers (IgM and B220) to identify B cells, 3) the isotype of the antibody produced (IgM, IgG1, IgG2a, or IgG2b) by a given cell, and 4) the affinity of that antibody for ovalbumin (Figure 5.15).

The data collected for each cell were compiled to construct a cellular profile of the humoral immune response observed for each mouse ([3]).

The graphical representation of the data makes it possible to assess visually the differences in

[3]pnas-fig5

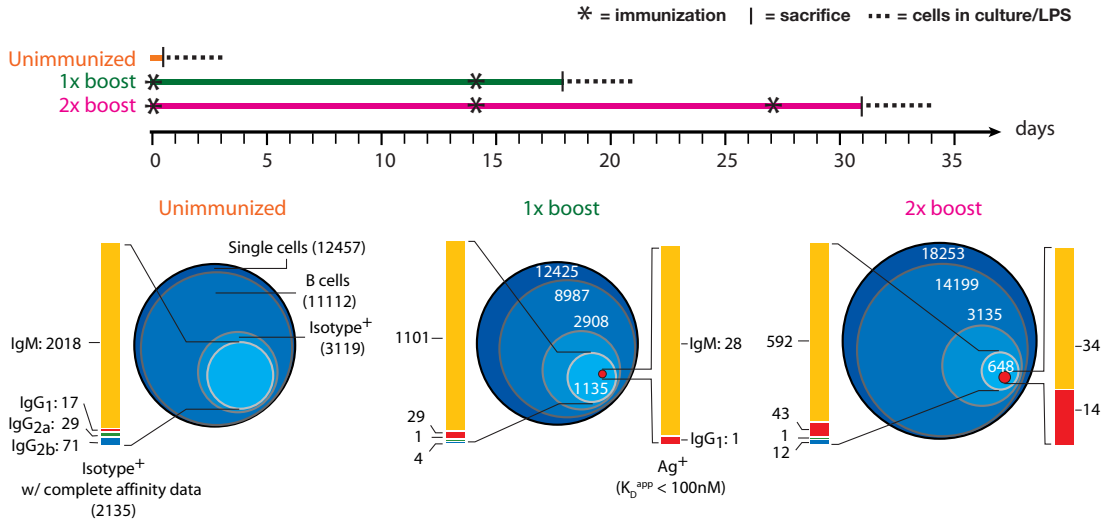

Figure 5.14: Profiles of antibody responses generated by immunized mice. (above) Immunization schedule used for the three mice profiled. (below) Graphical profiles of the populations of single cells characterized from each mouse. The total area of each circle is proportional to the number of cells enumerated with the phenotypes indicated. Single cells expressing either B220 or IgM on their surfaces were classified as B cells. The innermost blue-green circle represents the subset of individual secreting B cells for which a complete set of affinity data was obtained. The distribution of isotypes for this subset is indicated to the left of the colored circles for each mouse. Red circles represent the subset of individual antibody-secreting B cells classified as antigen-specific (with a dissociation constant, KD $<$ 100 nM). The distribution of isotypes in each of these subsets is shown to the right of the colored circles for each mouse. The quality of the microengraved microarrays can be estimated by comparing the number of single secreting B cells which yielded a complete set of affinity data to the total number of single secreting B cells detected (i.e., relative size of the green circle with respect to the enclosing isotype+ circle). Here, from left to right, 68.5%, 39% and 20.7%.

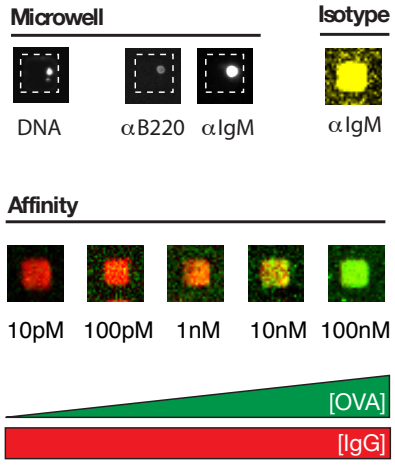

Figure 5.15: A representative set of the raw multivariate data collected by microengraving and immunofluorescence for primary splenocytes from the immunized mice.

the numbers of antibody-secreting cells and antigen-specific cells. For all three mice analyzed, B cells (B220+ and/or IgM+) comprised the major fraction of the individual cells isolated in microwells. Only 20–30% of those cells, however, actively secreted antibodies that were one of the four isotypes scored. Our analysis of which antibodies exhibited specificity and affinity for ovalbumin relied on this antibody-secreting population of B cells; we did not identify B cells that may have been specific for ovalbumin, but that did not secrete antibodies after treatment with LPS. We further restricted the analysis to those B cells for which a complete binding curve was available—that is, ones where high quality spots of the secreted antibody were present across all microarrays interrogated with ovalbumin. Either uneven fluorescent staining or a low ratio of signal-to-noise for a given element on one or more replicates excluded some cells from further consideration. These stringent conditions reduced the total size of each dataset, but data for a minimum of 648 individual antibody-secreting primary B cells from each mouse remained after this filtering.

Classification of antibodies generated in humoral immune responses.

The data collected for the immunized mice show, as expected, that primary immunization, followed by administration of a booster, yielded an increase in the percentage of antigen-specific B cells. To determine specificity for ovalbumin, we examined a plot of the binding curves that was rank-ordered according to their apparent dissociation constants, and then set a threshold at the point where the measured ratio failed to vary with concentration (Figure 5.16).

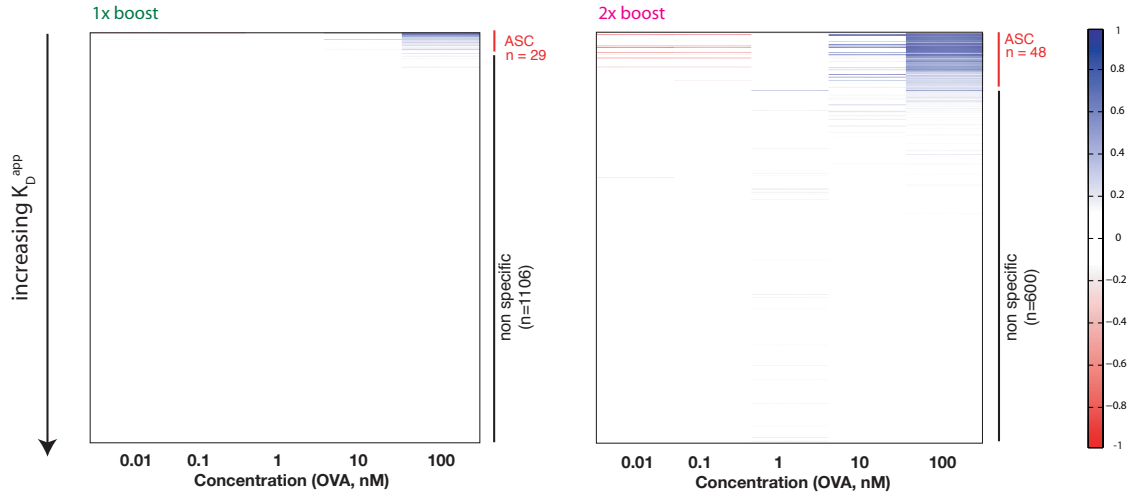

Figure 5.16: AffMaps for the populations of B cells from immunized mice for which data from all replicate microarrays were available (isotype and affinity). Each binding curve was fit and ordered according to the calculated apparent dissociation constant. The threshold for antigen-specific cells was set where the data threshold at the point where the measured ratio failed to vary with concentration.

This procedure excluded any B cells secreting antibodies with values of KD_appp greater than 100 nM. Following a single booster, 2.5% of the Ig secreting B cells identified in our analysis were antigen-specific (0.6% of the total population of splenocytes assayed). This percentage increased to 7.7% for a mouse receiving two boosters (1.3% of the total population of splenocytes assayed). The profiles for the three mice also show that affinity maturation and class-switching events occurred with repeated exposure, as expected. The frequency of class-switched antibodies within the antigen-specific population increased with the number of immunizations (from 3% IgG1 to 28%).

The median dissociation constant for the antigen-specific antibodies decreased with the number of immunizations (Figure 5.17).

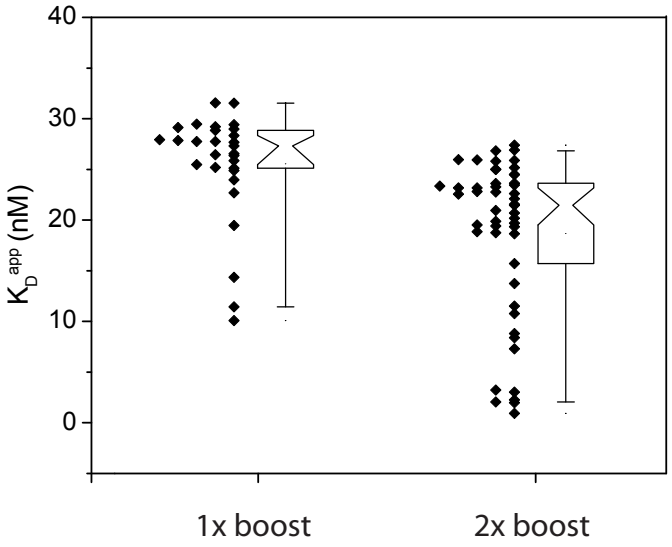

Figure 5.17: Plot of apparent dissociation constants calculated from individual binding curves for immunized mice. The box plots show the change in the median value of the measured apparent dissociation constants with increased immunization.

Analysis of the antigen-specific population of B cells by data clustering of the corresponding AffiMaps provided further insights to the diversity represented in those cells. To classify the relationships among the antigen-specific B cells identified in the immunized mice, we applied unsupervised hierarchical clustering to group cells based on both the isotype and binding curves of the antibodies they produced (Figure 5.19 and Figure 5.18).

The IgMs scored were largely indistinguishable by their apparent affinities (over the range of concentrations tested). The estimated values of KD_app for the IgG1 antibodies produced by the mouse that received three immunizations ranged from 1 nM to 25 nM; only seven cells produced antibodies with dissociation constants below 10 nM. Analysis by clustering organized these cells into five populations based on the shape of their binding curves. The ratio of cells producing antigen-reactive IgG to those producing IgM is smaller than one might expect based on the

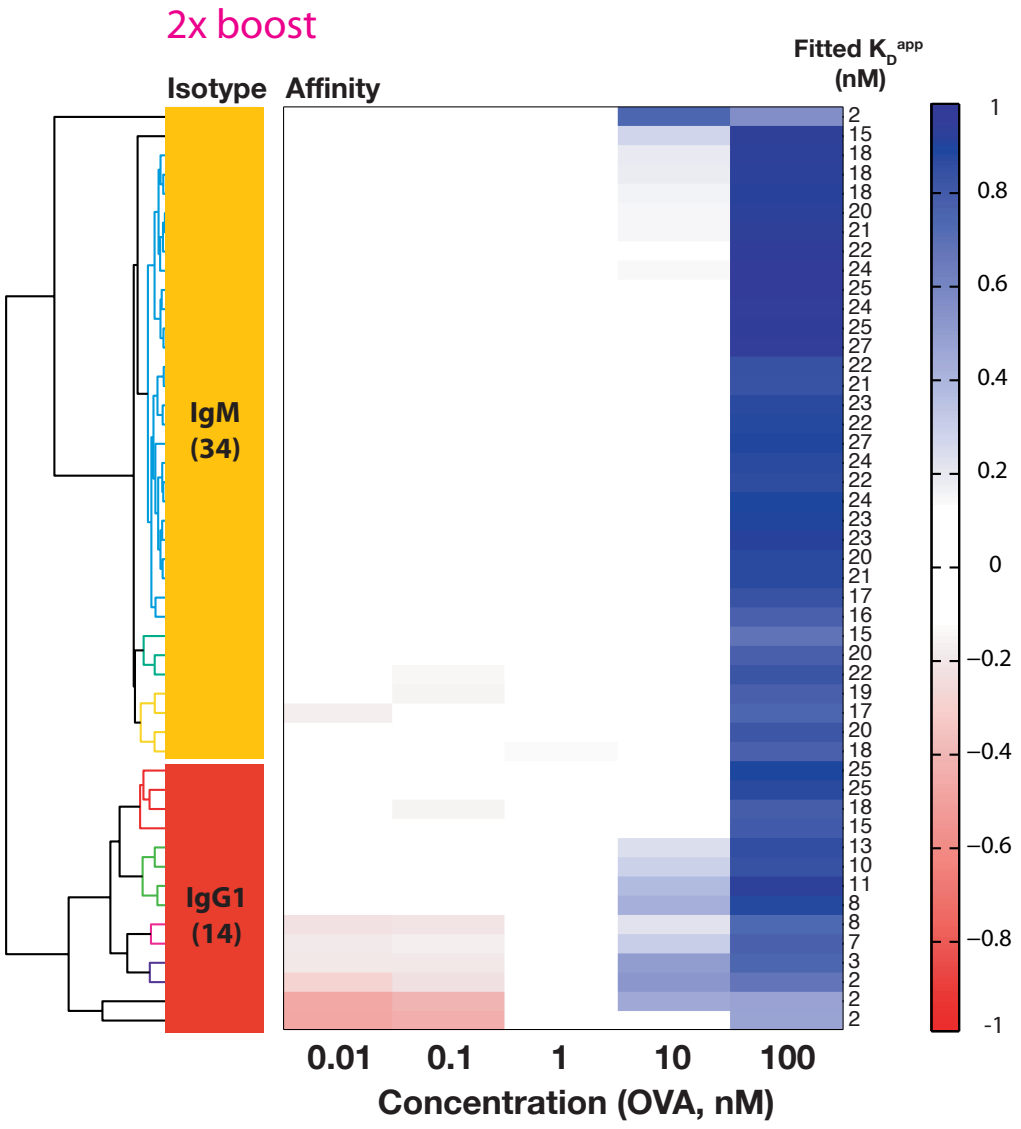

Figure 5.18: AffiMap of antigen-specific cells identified in a mouse after two boosters indicates the diversity in antibody response. Both the affinities and the isotypes were included in the hierarchical clustering (Euclidean/average), but separated in the figure for graphical clarity. The colored branches in the dendrograms indicate the cells that are most closely related. The apparent dissociation constants (nM) calculated from a langmuir curve fit for each clone are listed to the right of the data for each cell; these numerical values were not considered in clustering.

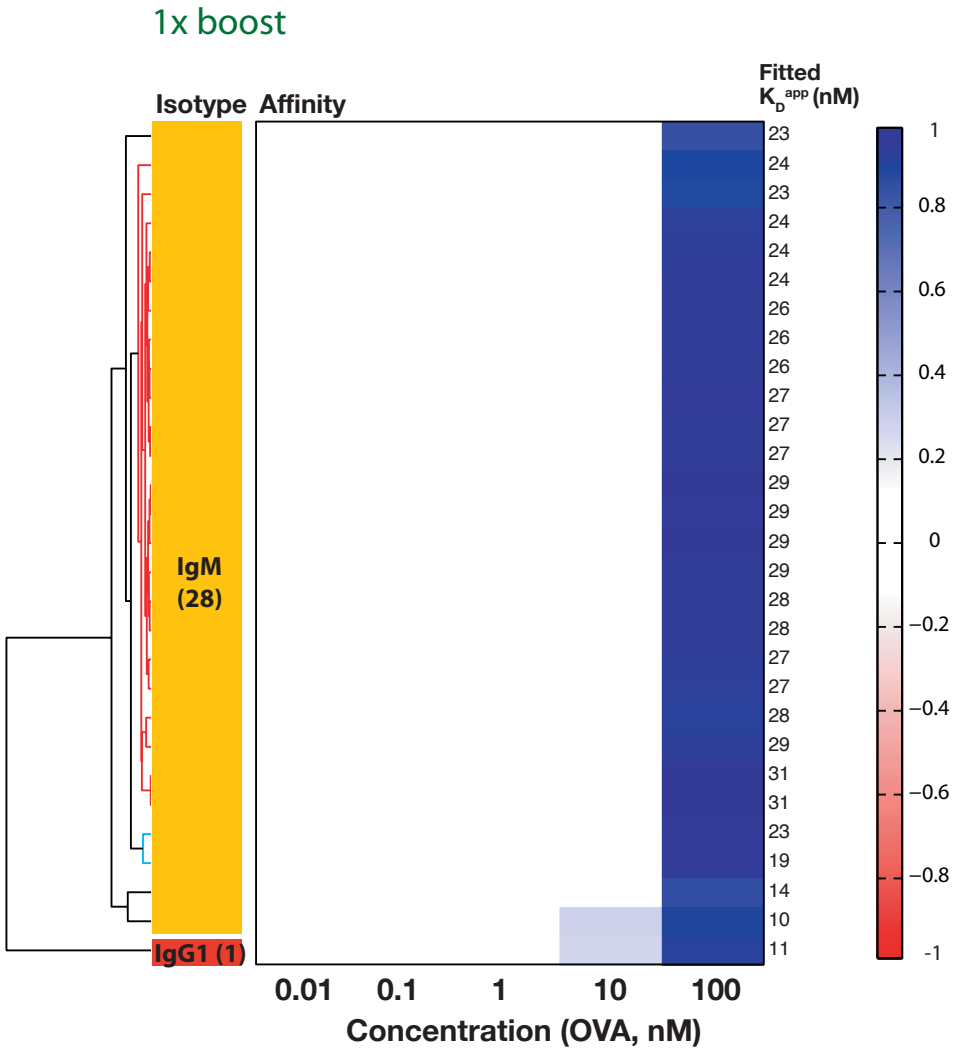

Figure 5.19: AffMap organized by hierarchical clustering (Euclidean/average) for immunized mouse receiving one booster. Both the affinities and the isotypes were included in the analysis by clustering, but set apart graphically for clarity. The colored branches in the dendrograms indicate closely related cells. The apparent dissociation constants (nM) calculated from a langmuir curve fit for each clone are listed to the right of the data for each cell; these numerical values were not considered in the clustering.

typical compositions of antigen-reactive sera from hyperimmunized animals where the principal component is often IgG [169]. One rationalization for this observation is that the population of cells taken from the spleen may include those IgM-producing B cells that are involved in a new primary response to the recurring antigenic stimulus. The short period of in vitro stimulation with LPS can also cause some B cells to proliferate more than others; this differential response may alter the frequency of antigen-specific cells that secrete antibodies compared to the frequency of splenic plasma cells present prior to stimulation ex vivo.

5.3 Discussion

We have developed a method to construct profiles that indicate the quality and diversity of a humoral immune response with single-cell resolution using data collected by both microengraving and immunofluorescence. The breadth of the correlated, quantitative data collected here for individual primary B cells has not been obtained, to the best of our knowledge, by any other single assay presently available. The data are equivalent to that acquired by flow cytometry (determination of lineages by surface-expressed markers on cells) and immunosorbant assays (isotype, specificity, and frequency), but with the advantage that these two sets of data are perfectly matched to each other. The data also include an estimate of the affinity of the antibody produced by a single B cell for an antigen of interest—a measurement that previously has required purified antibodies collected from expanded clonal populations of cells.

One significant advantage of the technique developed here over ELISpot or ELISA for estimating the affinity of an antibody is that the relative occupancy of the antibody is measured using antigen in solution. This approach minimizes confounding multivalent interactions of antibodies with plate-bound antigen or the relative capture of antigen on the surface in its native structure. Interactions from polyreactive, low-affinity antibodies can also induce false-positives. We do not know, however, the extent to which the format of the assay may overestimate KD_app for isotypes with different numbers of binding sites (e.g., IgG, IgM). This potential source of error makes the method best

suited for making relative comparisons among cells with the same isotypes. Furthermore, we have also found that applying antigens at concentrations at or above $\sim$0.5–1 μM results in significant non-specific staining that interferes with the quality of the data. This practical limit defines an upper bound on the values of KD_app that can be accurately estimated. The strength of the technique presented here is similar to that of other types of microarrays used for assessing gene expression: the method yields a diverse set of data with which to make relative comparisons among many individual events to determine patterns of events or unique outliers. For gene expression, measuring the differential expression of a gene under multiple conditions uniquely identifies genes affected in similar ways; here the binding curves generated from multiple arrays identify related populations of cells secreting a particular antibody. The cellular profiles constructed for the mice immunized with ovalbumin demonstrate how microengraving, in combination with immunofluorescence, can highlight those B cells actively contributing to a humoral immune response. The frequencies of antigen-specific B cells in the mice receiving immunizations reflect the subset of those cells that were actively secreting antibody after activation with LPS. Our estimated frequencies are similar to those determined previously by surface-staining and flow cytometry for mice undergoing a primary response to a hapten (nitrophenol) (1–2% of total splenocytes) [166]. The frequencies of antigen-specific B cells, however, can vary widely with the schedule of immunizations, the antigens, the genotypes of the mice, and the timing of sampling [165] [166] [166]. We were able to capture antibodies secreted by splenocytes taken directly from a mouse (data not shown), but the quality of the data (signal-to-noise) was insufficient for robust quantitative analysis. We anticipate that technical refinements to microengraving will improve the sensitivity as well as the efficiency of capture, and make mitogenic stimulation unnecessary. We believe that further development of the approach described here will enable rapid monitoring of humoral immune responses in humans to vaccines. Such measures could facilitate the rational design of vaccines by providing insight to the diversity and maturation of antibody-producing clones induced by a candidate vaccine among different populations of patients as a function of formulation or adjuvants [170]. This detailed knowledge could potentially reduce the cost of vaccine trials by improving the criteria for inclu-

sion in the study, and would complement related genomic and proteomic analyses. In addition to vaccine development, the approach here also should improve clinical research on the pathogenesis of both infectious and autoimmune diseases [171] [172] [173] [174] [175] [176] [177].

5.4 Methods

Cell culture.

Hybridoma cell line 099–01, secreting anti-OVA was obtained from Statens Serum Institut. The Y3 cell line, secreting anti-H–2Kb antibody was a kind gift of Dr. P Cresswell. Anti-H–2Kb hybridoma lines (c127 and c136) were described previously (9). All hybridomas were cultured in DMEM (Gibco) supplemented with 10% FBS, 50 units penicillin/50 µg streptomycin, 20 mM HEPES, 50 µM 2-mercaptoethanol, 1 mM sodium pyruvate and 0.1 mM nonessential amino acids (Gibco). Cells were maintained at 5% CO2, at 37° C, and split every 2–3 days.

Fluorescent reagents.

The following reagents were purchased from Invitrogen and used as received: Ovalbumin Alexa Fluor 555 conjugate (O–34782); Alexa Fluor 647 goat anti-mouse IgG1 (A21240); Alexa Fluor 594 goat anti-mouse IgM (A21044); Alexa Fluor 555 goat anti-mouse IgG2a (A21137); Alexa Fluor 488 goat anti-mouse IgG2b (A21141); Alexa Fluor 647 goat anti-mouse IgG (H+L) (A21235), CellTracker Blue CMAC (C2110), CellTracker Red CMTPX (C34552), CFSE (C34554).

Labeling hybridomas for imaging.

To label hybridomas with one of the three cytosolic dyes used (CellTracker Blue, CellTracker Red, or CFSE), cells were incubated in serum-free media containing one of the dyes (10 µM) for 30 min at 37 ºC. The cells were pelleted, and the media replaced with DMEM/10% FBS. The cells were incubated for 30 min at 37 ºC, and then used for microengraving.

Mice.

For immunizations, a hybrid mouse strain between 129 and Balb/c was used. These mice also contain the XBP–1 gene flanked by lox-P sites, but expression of the XBP–1 gene is normal, as determined by biochemical analysis [courtesy of Dr. Laurie Glimcher; C. A. Hu, data unpublished].

Immunizations and splenocyte stimulation.

Mice were immunized intraperitoneally with a mixture of OVA (50 μg) and HEL (50 μg) as an emulsion with complete Freund's adjuvant (Sigma), and boosted at day 14, and day 27 with an emulsion of antigen and incomplete Freund's adjuvant. Splenocytes were harvested from the mice four days after their last booster. The splenocytes were harvested by disaggregation of the spleen, and the resulting suspensions of splenocytes (1 106 in RPMI 1640) were stimulated with lipopolysaccharide (LPS; 20 μg/ml) for three days.

Microengraving.

The microengraving system was employed as described. [167] Briefly, a prepolymer of PDMS (Sylgard 184, Dow Corning) was cast against molds of SU–8 posts manufactured by photolithography on 3 inch silicon wafers (Silicon Sense, Nashua, NH). The microwells (50 μm 50 μm 50 μm) were arranged in 40x40 blocks on a pitch of 100 μm; the blocks were arranged in a 4 13 grid. After curing for 12–24 h, the arrays of microwells were removed, treated in an oxygen plasma for 60 s, and placed into a solution of bovine serum albumin (0.5% w/v) for at least 1 h at room temperature. The molds were gently rinsed three times with phosphate-buffered saline, and then cells were deposited into the wells by gravity. Typically, 0.5 mL of a suspension of cells (5105 cells/mL) were added to the stamp for 3–5 min; the degree of loading was observed qualitatively under a microscope. Excess cells were lightly rinsed from the surface, and then excess media was aspirated from the surface. The molds were placed into contact with glass slides (Super-Epoxy 2™, Telechem Intl.) pre-coated with goat anti-mouse IgG capture antibody (Southern Biotech 1030–01) or goat anti-

mouse IgG (H+L) (1010–01), and blocked with BSA. (Capture antibodies were deposited from a semi-optimized coating buffer (50 mM Borate, pH 9, 80 mM Sucrose, 50 mM NaCl) (J. Ronan et al., unpublished data).) The stamp supported on the glass slide was placed in a hybridization chamber (DT–1001 with set of screws used for DT–2002, Die-Tech, San Jose, CA) to hold the assembly together, and incubated for 30 min at 37 deg C. After incubation, stamp was gently removed from the glass slide, and placed immediately into a reservoir of media. The glass slide was placed in a blocking buffer (1% BSA/0.05% Tween 20/PBS) for 1 h at room temperature. For replicate prints, the process above was repeated with the same stamp and a pre-coated slide. The slides were analyzed by incubating with fluorescent antibodies and/or proteins of interest as previously described [167].

Hybridoma experiments

Cells from three monoclonal hybridomas were independently labelled with cytosolic dyes (c136 with CellTracker Blue CMAC, c127 with CellTracker Red CMTPX, Y3 with CFSE) (Invitrogen), and then deposited into microwells. Seven replicate microarrays were generated by microengraving, and probed with anti-mouse Ig (H+L) (Alexa 647, 10 nM, Invitrogen) and tetrameric H–2Kb-streptavidin (Alexa 532, 1 pM to 500 nM). After engraving, the cells were labelled with a nuclear stain (Hoescht 33342, Invitrogen), and then imaged by epifluorescence on an automated, inverted microscope.

Immune response profiling

Splenocytes were harvested from a hybrid strain of mice (between 129 and Balb/c) at days 0, 15 and 30 following immunization and stimulated for three days with LPS (20 μg/ml). The cells were then deposited into microwells, and six replicate microarrays were generated by microengraving. After engraving, the cells were fixed with 4% paraformaldehyde, labelled with a nuclear stain (Hoescht 33342), anti-IgM (Alexa 647), and anti-B220 (rhodamine), and imaged. One microar-

ray was stained with isotype-specific antibodies: anti-IgM (Alexa 594), anti-IgG1 (Alexa 647), anti-IgG2a (Alexa 532) and anti-IgG2b (Alexa 488). Five replicates were probed with anti-mouse Ig (H+L) (Alexa 647) (10 nM) and increasing concentrations of ovalbumin (Alexa 555) (10 pM to 100 nM).

Data analysis

Fluorescence images of the cell-loaded PDMS device were mined by custom routines in Metamorph software (Molecular Devices). Glass slides supporting the printed microarrays were imaged on a microarray scanner (Genepix 4000B, 4200AL) and mined by Genepix Pro software (Molecular Devices). Multidimensional data was correlated in MATLAB (The MathWorks, Inc). All subsequent data filtering, analysis and clustering were performed by custom code written in MATLAB (available upon request). Additional methods. Additional details for all experimental methods are available in the Supplementary Material Descriptions, including cell culture, fluorescent reagents, mice and immunization methods.

Immunofluorescence.

For analysis of cells in the wells of the PDMS grid, cells were fixed in 4% paraformaldehyde for 15 min at room temperature, then rinsed, and stained using Alexa Fluor 568-labeled goat anti-mouse IgM (μ chain) antibody, Alexa Fluor 488-labeled anti-mouse CD45R/B220 antibody (clone RA3–6B2), and Hoechst 33342 (1 μg/ml). Arrays of microwells were covered with a thin layer of PBS and placed face down onto the surface of a clean glass slide. All images were acquired on an inverted epifluorescence microscope (Nikon Eclipse TE2000-E) equipped with a Hamamatsu Orca AG camera, and an automated x-y translation stage (Prior Instruments). Overlapping images were acquired for the entire array using the Scan Slide module in the MetaMorph software package (v7.1, Molecular Devices, Sunnyvale, CA). Images of complete blocks of microwells (40 40 wells 52 blocks per array) were assembled using the Scan Slide module in MetaMorph for each data channel acquired.

Semi-automated tabulation of the number of cells found in each well, and their expressed surface markers, was generated using custom macros written in MetaMorph (available upon request).

Microarray imaging.

For analysis of slides probed with fluorescent dye reagents, we used microarray scanners (Genepix 4000B and 4200AL) to collect the images of the microarrays in either two (affinity slides) or four channels (isotype slides). The accompanying software (Genepix Pro, Molecular Devices) was used to extract features and fluorescent data. Raw numerical data was then exported to MATLAB® (The MathWorks, Natick, MA) for multidimensional correlation, filtering and statistical analysis.

Data correlation and analysis.

Raw numeric data obtained from isotype and affinity slides were imported into MATLAB®. Further processing was performed by custom MATLAB® scripts and open source subroutines (available upon request). Scanned spots with poor signal quality (high spot covariance, low signal to noise ratio and/or high saturation levels) were filtered from the dataset. Remaining good quality spots were then ranked as Ig+, by setting a threshold on the background corrected median intensity, total intensity and signal to noise ratio in the corresponding channel. The isotype slides were taken as the reference measure of the Ig-secreting cells in a given array. The first print generated with each set of microwells was used to evaluate isotypes, and these prints consistently were of higher quality (higher signal to noise ratio, more positive spots) than the scans obtained for the affinity slides. Filtered information from a) isotype slides, b) affinity slides and c) cellular stains (imaged by epifluorescence microscopy) was correlated on the basis of relative spatial coordinates (column, row, block number). Once the full multidimensional dataset was assembled, a subset of wells was selected that contained single cells and yielded Ig+ elements in all slides (1 isotype plus 5 affinity for the mouse dataset; 7 affinity and no isotype for the anti-Kb dataset). Unsupervised clustering, either by hierarchical or partitioning methods, was performed in MATLAB® to

resolve distinct classes of clones. Cophenet distances and coefficients were computed to assess the quality of different hierarchical methods. Only hierarchical clustering methods yielding an average cophenet coefficient $> 90\%$ were employed. Silhouette and gap statistics were similarly employed to determine the adequate number of clusters used in partitioning methods (i.e. k-means). Principal components analysis was used routinely to monitor the distribution of the selected dataset and to control the quality of the clustering results.

Determination of antigen-specific clones in the mouse dataset.

Each vector of data corresponding to a single cell was fit with a Langmuir binding isotherm by non linear least squares (95% confidence interval). The affinity curves obtained were sorted in ascending order, based on the fitted apparent dissociation constant Kdapp. A threshold was determined following visual inspection of the data to facilitate the identification of clones with measurable Kdapp (< 100 nM). The frequency of antigen-specific clones was calculated with respect to the total number of wells shown to be consistently secreting in the isotype slide. Isotype information was obtained for each antigen-specific clone.

Evaluation of accuracy of clustering applied to model hybridoma data.

The identity of each cell was determined by analysis of the immunofluorescence data collected for the cells with cytosolic stains in the microwells (Figure 5.10). Each unsupervised method (hierarchical or k-means) was compared to the cell marker count (assumed as representing the true classification) and a value for the sensitivity (true positives over combined true positives and false negatives), specificity (true negatives over combined true negatives and false positives), positive predictive value (PPV) and negative predictive value (NPV) were obtained for each of the three clonal classifications. The average across the three clones is included in Table 5.1.

Part IV

Discussion

In the previous chapters I have described a series of tools for the high-throughput and high resolution characterization of both the immune system and the human microbiome. In chapter 5 I present uEn, a technique to profile the antibody-mediated cell response and shown its application toward the study of vaccinations. In chapter 4, I introduce the use of microfabrication to perform "pulse-chase" experiments on single lymphocytes with physiologically relevant ligand doses. The remaining work Part II, introduces the use of supervised learning in the analysis of newly available microbiome data and demonstrates its applicability toward differentiating healthy and diseased pediatric IBD patients on the basis of their stool microbiome.

The integration of these and other [178] systems-level tools will allow us to define the relationship between host immune system and microbiome and hopefully elucidate the molecular mechanisms responsible for this delicate and vital balance. What factors trigger changes in gut inflammation? Are these factors microbial, environmental or host-derived? What predisposes and precipitates the pathogenic behaviour of *pathobionts*?

By investigating immune system status as a whole with rapid analysis of shifts in the microbiota, one can begin to answer some of these questions. The resulting multidimensional data will identify key associations and improve the resolution of longitudinal observations. Discovering significant correlations and dynamic shifts in either microbiota or immune system will, in turn, direct further biochemical investigations aimed at defining the underlying biological mechanisms. Before each of the methods presented in this work can be effectively integrated with the others, specific developments and technical improvements are necessary.

Microengraving for instance, though a relatively simple experimental procedure, still requires some amount of training and suffers from poor automation. I've participated in early efforts towards the optimization and standardization of the technique, conducted in collaboration with [179, J.Ronan] and [180, A.Oguniunny]. Since then substantial technological development has taken place [181] [182] [183] and has been directed toward i) further standardization of the microengraving protocol, ii) the creation of specific devices and engineering solutions that simplify the current process and iii) the streamlining of data mining instrumentation and algorithms.

The establishment of standard protocols and a set of appropriate tools [180] has improved the quality and repeatability of the microengraving process. Staining of glass slides is done by means of commercially available automation to ensure uniform signal. Chip fabrication is standardized with the use of an injection mold to increase the speed of production. Lastly, the placement of the chip in contact with the glass slide is performed by means of an appropriate mechanical device instead of relying on manual placement. Data mining has also been streamlined: improvement on chip design allows automatic alignment of images and identification of wells; the use of dedicated imaging instrument significantly increases the speed of data acquisition; and the development of custom software for image analysis and data correlation removed the largest bottleneck in the pipeline.

These efforts have already enabled the routine collection of data by microengraving [184], adding the ability to select relevant antibody producing cells and profiling the status of humoral immune responses to the panel of techniques at the disposal of the immunologist. Thus it is now possible to determine the impact different microbial communities will have on the dynamic and the secretions of lymphocyte populations, by applying microbial factors or entire microbial populations to the microengraving chip (or extracting mucosal lymphocytes at different stages of gut colonization). Changes in secretory patterns undetectable by traditional methods may be diagnostic of disease initiation or, at least, informative about its progression.

For this specific reason, other secreted factors will be of interest when profiling the dynamics of immune responses, including cytokines, chemokines, ROS as well as pathogen derived secretions, just to name a few. Cytokines in particular mediate lymphocyte activity and homeostasis, as well as the regulation of dendritic cells and antigen presenting cells in general. By engineering improvements in the signal-to-noise ratio for the printed protein microarrays and by expressing cytokine-specific capture antibodies, uEn has been expanded by [183] to include the analysis of cytokine release.

Cytokine measurements, in turn, allow for the detection and profiling of other fundamental players of the immune response in the gut mucosa, such as T lymphocytes (Tregs and Th17) and dendritic cells. Since it is well known that the equilibrium between Tregs and Th17 cells in the gut

determines inflammatory state, it may be possible to develop an assay for inflammation or, better, individuate the factors responsible for the initiation of the inflammatory process.

Technological developments have also permitted routine recovery of the single cells interrogated [181], permitting the study and characterization of rare lymphocytes [185]. This ability will be of paramount importance when tools will be applied to the study of mucosal immunity, since lymphocyte populations in the mucosa are less abundant and harder to isolate than those routinely extracted from peripheral blood. It will be possible to recover individual lymphocytes demonstrating activity upon interaction with microbial factors and thus further characterize them molecularly and genetically.

Lastly, no quantitative models exist to describe the reaction of the immune system to microbiota. Further technological advances in microengraving will allow the acquisition of a greater number of signals, thus increasing the number of variables measurable at each stage of the process. In turn, this will enable the formation of more detailed theoretical models, potentially capable of predicting the course of autoimmune disease and infection.

Exposure of single immune cells to physiologically appropriate quantities of microbial factors can be useful for defining molecular details of the interaction between single immune cells and the microbiota. The tool demonstrated for the **mapping of surface dynamics** of single lymphocytes can easily be adapted to this task, by either employing microbial factors as ligands or by enabling contact between microbes and single antigen-presenting cells.

Development and redesign of the method shown in this work will however be necessary to surmount its principal limitation: low throughput. The construction of parallel array (or chamber microchips) is a practical improvement and has already been demonstrated for different applications. [186] [187] [188]Technological advances in fluorescence microscopy are also needed in order to image larger fields of view with the same level of resolution shown in this work. [189] [190] [191]

The validity of this approach, however, has been amply demonstrated. Similar types of high-resolution single cells experiments have been devised through microfabrication and high-resolution microscopy. [192] Tolar et al [150] and Sohn et al. [149] have observed the changes occurring dur-

ing the initiation of B cell signalling and the microclustering requirements on the surface of B cells by means of fluorescence energy transfer measurements. Mapping of the response of immune cells to chemotactic stimuli similar to the ones found in vivo has also been conducted with the aid of microfabricated chips. [155] [193]

The biological insight granted by these high-resolution techniques promises to elucidate many details regarding the intracellular pathways and surface interactions of single immune cells with their microenvironment and at the immunological synapse. [194] The application of these technologies to antigen presenting cells and lymphocytes in the gut mucosa, will allow us to interrogate the molecular pathways responsible for microbial recognition and processing.

Development of high throughput methods to interrogate immunity by definition requires sophisticated analysis algorithms to evaluate the resulting data output. Computational tools developed for the analysis of **microbiome composition** will be highly relevant in the characterization of mucosal immunity, both in the gut and elsewhere in the body. Machine learning tools reduce the complexity of monitoring large bacterial communities and are sensitive to subtle shifts in microbial composition. Changes in the parameters measured for the immune system can then be correlated with these shifts with the aim of determining causative factors, or at least key biomarkers. Information about the host immune system can also be integrated by performing classification of taxonomical data against any of the parameters measured while characterizing the immune response. Machine learning models will then identify the significant changes in the microbiota that accompany variation in immune parameters.

Future work will include the evaluation of other existing machine learning algorithms and feature selection strategies. Although a limited number of models have been discussed herein and ensemble methods such as RandomForest proved optimal, novel algorithms are actively being developed in the machine learning research community. In particular, efforts are focusing on dealing with larger datasets in parallel and the adaptation of ensemble models to data with large number of features [195]. These are two kind of advances that would greatly benefit classification of microbial sequencing data.

One potential method for the unbiased evaluation of algorithms is the organization of a machine learning competition on microbiome data [132]. The hope is to stimulate the development of new algorithms to classify microbial community data, by providing an anonymized dataset of patient sequences and metadata on an open-access platform. It will also be important to assess whether the inclusion of phylogenetic information (or ecological information) would positively affect the accuracy of supervised learning and whether it can provide further insight in the community structure at a specific site. Even though classic machine learning algorithms could be applied directly to raw microbiome data, it is arguable that better performance in classification would be achieved by incorporating microbiological and evolutionary knowledge in a classification algorithm. Correct preprocessing and representation of the input data generally improves the performance of machine learning algorithms, often requiring less input data to achieve the same classification accuracy. Previous knowledge about the input data or its underlying structure can be integrated into classification algorithms, and varying input representation can highlight different patterns in the data. In the case of microbiome data, it will be necessary to evaluate whether the inclusion of phylogenetic information can improve accuracy and relax training dataset requirements. Any improvement in accuracy can improve the rate and the quality at which associations between microbial composition and disease are found.

The integration of machine learning algorithms into sequencing processing pipelines is also a future direction for the work illustrated here. Supervised learning methods can prove to be a powerful tool in evaluating the relative importance of various sequencing parameters (sequencing platform, sequencing depth, read length, variation between centers,etc.). The rapid comparison of sequencing results coming from control experiments can be a direct and straightforward application for approaches based on learning.

Perhaps most importantly, it will be crucial to apply and adapt the tools presented here to the analysis of longitudinal studies. Key to the advance of our understanding of host-microbiota interaction will be the collection of longitudinal data on microbial composition and immune status. Only by observing the transition from normal physiology to inflammatory state and its sequela, will

it be possible to infer the correct trigger and the order of each step in the cascade leading to disease. Software and machine learning algorithms will have to be adapted to deal with time-varying data, by allowing for the real-time updating of training sets. It will also be necessary to develop an appropriate structure for the input data that takes into account time as an independent variable so that it can be parsed by machine learning algorithms. As the algorithms would be dealing with time series data, it may be more appropriate to build models for the trend of parameters rather than their discrete values.

Finally, a focus of future work will be the integration of ecological theory in the study of microbial populations living in association with the human host. Understanding the biology of the microbiome and how it interacts with the host immune system will require a development of ecological theory beyond the current levels. Created to describe ecology and evolution of animal species, current ecological theory does not suffice to describe and predict the ecology of microbial species. A clear definition of species in microorganisms is lacking. Mobile genetic elements make this definition even more challenging. Moreover, despite the recent improvements brought about by 16S sequencing, MLST, and pyrosequencing, measuring the richness and spatial distribution of microbial species remains an imprecise quantitative exercise. These are challenges that will have to be surmounted to analyze and predict the impact of microbes on their host, as well as their spatiotemporal dynamics. [196] Formal construction of theoretical hypotheses is arguably the best way to tackle these challenges and advance the field of microbiome biology. Theoretical formulations will be able to generate predictions, which then can be experimentally validated by the set of new tools being created.

The type of analysis described will allow us to find more quickly and consistently important patterns in microbiome data, a desirable outcome in a field where data is currently being accumulated at an impressive rate and where there exists clear clinical correlations of immediate impact.

Ultimately, the various tools developed as part of this work will be integrated to provide a unified view of the host immune system and the microbiota. The resulting picture will hopefully deepen our knowledge of the complex and deeply interesting interaction humans maintain with

the microbes around us.

Bibliography

[1] Ruth E Ley, Fredrik Bäckhed, Peter Turnbaugh, Catherine a Lozupone, Robin D Knight, and Jeffrey I Gordon. Obesity alters gut microbial ecology. *Proc Natl Acad Sci USA*, 102(31):11070–11075, August 2005. ISSN 0027-8424. doi: 10.1073/pnas.0504978102. 10

[2] Ruth E Ley, Daniel A Peterson, and Jeffrey I Gordon. Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine. *Cell*, 124(4):837–848, February 2006. doi: 10.1016/j.cell.2006.02.017. 10, 14, 69

[3] Peter J Turnbaugh, Fredrik Bäckhed, Lucinda Fulton, and Jeffrey I Gordon. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe*, 3(4):213–223, April 2008. doi: 10.1016/j.chom.2008.02.015. 10

[4] Peter J Turnbaugh, Ruth E Ley, Michael A Mahowald, Vincent Magrini, Elaine R Mardis, and Jeffrey I Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122): 1027–1031, December 2006. doi: 10.1038/nature05414. 10

[5] R Balfour Sartor and Marcus Muehlbauer. Microbial host interactions in IBD: Implications for pathogenesis and therapy. *Curr Gastroenterol Rep*, 9(6):497–507, December 2007. doi: 10.1007/s11894-007-0066-4. 10, 17, 24

[6] Charles O Elson, Yingzi Cong, Vance J McCracken, Reed A Dimmitt, Robin G Lorenz, and Casey T Weaver. Experimental models of inflammatory bowel disease reveal innate, adaptive, and regulatory mechanisms of host dialogue with the microbiota. *Immunol Rev*, 206:260–276, August 2005. doi: 10.1111/j.0105-2896.2005.00291.x. 10

[7] D N Frank, A L St. Amand, R A Feldman, E C Boedeker, N Harpaz, and N R Pace. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences*, 104(34):13780–13785, August 2007. doi: 10.1073/pnas.0706625104. 10, 17, 25, 26, 31, 34, 38, 52, 53, 73

[8] Onyinye I Iweala and Cathryn R Nagler. Immune privilege in the gut: the establishment and maintenance of non-responsiveness to dietary antigens and commensal flora. *Immunol Rev*, 213:82–100, October 2006. doi: 10.1111/j.1600-065X.2006.00431.x. 10, 13

[9] Mairi C Noverr and Gary B Huffnagle. Does the microbiota regulate immune responses outside the gut? *Trends in microbiology*, 12(12):562–568, December 2004. ISSN 0966-842X. doi: 10.1016/j.tim.2004.10.008. 10

[10] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H Bjø rn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Sø ren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, Peer Bork, S Dusko Ehrlich, and Jun Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, March 2010. ISSN 1476-4687. doi: 10.1038/nature08821. 11

[11] Yun Kyung Lee and Sarkis K. Mazmanian. Has the microbiota played a critical role in the evolution of the adaptive immune system? *Science (New York, N.Y.)*, 330(6012):1768–73, December 2010. ISSN 1095-9203. doi: 10.1126/science.1195568. 11, 12, 13, 16

[12] Nadine Cerf-Bensussan and Valérie Gaboriau-Routhiau. The immune system and the gut microbiota: friends or foes? *Nature reviews. Immunology*, 10(10):735–44, October 2010. ISSN 1474-1741. doi: 10.1038/nri2850. 11, 12, 16

[13] Carl J. Yeoman, Nicholas Chia, Suleyman Yildirim, Margret E. Berg Miller, Angela Kent, Rebecca Stumpf, Steven R. Leigh, Karen E. Nelson, Bryan a. White, and Brenda a. Wilson. Towards an Evolutionary Model of Animal-Associated Microbiomes. *Entropy*, 13(3):570–594, February 2011. ISSN 1099-4300. doi: 10.3390/e13030570. 12

[14] Paul Wilmes, Sheri L Simmons, Vincent J Denef, and Jillian F Banfield. The dynamic genetic repertoire of microbial communities. *FEMS microbiology reviews*, 33(1):109–32, January 2009. ISSN 0168-6445. doi: 10.1111/j.1574-6976.2008.00144.x. 12

[15] Charles J Goodnight. Evolution in metacommunities. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 366(1569):1401–9, May 2011. ISSN 1471-2970. doi: 10.1098/rstb.2010.0290. 12

[16] L. V. Hooper. Commensal Host-Bacterial Relationships in the Gut. *Science*, 292(5519):1115–1118, May 2001. ISSN 00368075. doi: 10.1126/science.1058709. 12

[17] Sarkis K Mazmanian, June L Round, and Dennis L Kasper. A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature*, 453(7195):620, May 2008. doi: doi:10.1038/nature07008. 12, 13, 16

[18] June L Round and Sarkis K Mazmanian. The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews Immunology*, 9(5):313, May 2009. ISSN 1474-1741. doi: doi:10.1038/nri2515. 13, 16

[19] Shaoguang Wu, Ki-Jong Rhee, Emilia Albesiano, Shervin Rabizadeh, Xinqun Wu, Hung-Rong Yen, David L Huso, Frederick L Brancati, Elizabeth Wick, Florencia McAllister, Franck Housseau, Drew M Pardoll, and Cynthia L Sears. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nature medicine*, 15(9):1016–22, September 2009. ISSN 1546-170X. doi: 10.1038/nm.2015. 13

[20] Arlette Darfeuille-Michaud, Jérôme Boudeau, Philippe Bulois, Christel Neut, Anne-Lise Glasser, Nicolas Barnich, Marie-Agnès Bringer, Alexander Swidsinski, Laurent Beaugerie, and Jean-Frédéric Colombel. High

prevalence of adherent-invasive Escherichia coli associated with ileal mucosa in Crohn's disease. *Gastroenterology*, 127(2):412–21, August 2004. ISSN 0016-5085. 13

[21] Marcello Chieppa, Maria Rescigno, Alex Y C Huang, and Ronald N Germain. Dynamic imaging of dendritic cell extension into the small bowel lumen in response to epithelial cell TLR engagement. *The Journal of experimental medicine*, 203(13):2841–52, December 2006. ISSN 0022-1007. doi: 10.1084/jem.20061884. 14

[22] Olga Schulz, Elin Jaensson, Emma K Persson, Xiaosun Liu, Tim Worbs, William W Agace, and Oliver Pabst. Intestinal CD103+, but not CX3CR1+, antigen sampling cells migrate in lymph and serve classical dendritic cell functions. *The Journal of experimental medicine*, 206(13):3101–14, December 2009. ISSN 1540-9538. doi: 10.1084/jem.20091925. 14

[23] G Eberl and M Lochner. The development of intestinal lymphoid tissues at the interface of self and microbiota. *Mucosal immunology*, 2(6):478–85, November 2009. ISSN 1935-3456. doi: 10.1038/mi.2009.114. 14

[24] Wendy S Garrett, Carey A Gallini, Tanya Yatsunenko, Monia Michaud, Andrea Dubois, Mary L Delaney, Shivesh Punit, Maria Karlsson, Lynn Bry, Jonathan N Glickman, Jeffrey I Gordon, Andrew B Onderdonk, and Laurie H Glimcher. Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. *Cell host & microbe*, 8(3):292–300, September 2010. ISSN 1934-6069. doi: 10.1016/j.chom.2010.08.004. 14, 16

[25] Lora V Hooper and Andrew J Macpherson. Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nature reviews. Immunology*, 10(3):159–69, March 2010. ISSN 1474-1741. doi: 10.1038/nri2710. 14

[26] David Artis. Epithelial-cell recognition of commensal bacteria and maintenance of immune homeostasis in the gut. *Nature Reviews Immunology*, 8(6):411, June 2008. doi: doi:10.1038/nri2316. 14

[27] Les Dethlefsen, Margaret McFall-Ngai, and David A Relman. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature*, 449(7164):811–8, October 2007. ISSN 1476-4687. doi: 10.1038/nature06245. 14

[28] Andrew J Macpherson and Nicola L Harris. Interactions between commensal intestinal bacteria and the immune system. *Nature reviews. Immunology*, 4(6):478–85, June 2004. ISSN 1474-1733. doi: 10.1038/nri1373. 15

[29] Laurent Peyrin-Biroulet, Julia Beisner, Guoxing Wang, Sabine Nuding, Sajit Thottathil Oommen, Denise Kelly, Erika Parmentier-Decrucq, Rodrigue Dessein, Emilie Merour, Philipe Chavatte, Teddy Grandjean, Aude Bressenot, Pierre Desreumaux, Jean-Frédéric Colombel, Béatrice Desvergne, Eduard F Stange, Jan Wehkamp, and Mathias Chamaillard. Peroxisome proliferator-activated receptor gamma activation is required for maintenance of innate antimicrobial immunity in the colon. *Proceedings of the National Academy of Sciences of the United States of America*, 107(19):8772–7, May 2010. ISSN 1091-6490. doi: 10.1073/pnas.0905745107. 15

[30] Heather L Cash, Cecilia V Whitham, Cassie L Behrendt, and Lora V Hooper. Symbiotic bacteria direct expression of an intestinal bactericidal lectin. *Science (New York, N.Y.)*, 313(5790):1126–30, August 2006. ISSN 1095-9203. doi: 10.1126/science.1127119. 15

[31] Shipra Vaishnava, Cassie L Behrendt, Anisa S Ismail, Lars Eckmann, and Lora V Hooper. Paneth cells directly sense gut commensals and maintain homeostasis at the intestinal host-microbial interface. *Proceedings of the National Academy of Sciences of the United States of America*, 105(52):20858–63, December 2008. ISSN 1091-6490. doi: 10.1073/pnas.0808723105. 15

[32] Masayuki Tsuji, Noriko Komatsu, Shimpei Kawamoto, Keiichiro Suzuki, Osami Kanagawa, Tasuku Honjo, Shohei Hori, and Sidonia Fagarasan. Preferential generation of follicular B helper T cells from Foxp3+ T cells in gut Peyer's patches. *Science (New York, N.Y.)*, 323(5920):1488–92, March 2009. ISSN 1095-9203. doi: 10.1126/science.1169152. 15

[33] Andrea Cerutti and Maria Rescigno. The biology of intestinal immunoglobulin A responses. *Immunity*, 28 (6):740–50, June 2008. ISSN 1097-4180. doi: 10.1016/j.immuni.2008.05.001. 15

[34] Seth Rakoff-Nahoum, Justin Paglino, Fatima Eslami-Varzaneh, Stephen Edberg, and Ruslan Medzhitov. Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell*, 118 (2):229–41, July 2004. doi: 10.1016/j.cell.2004.07.002. 15

[35] Koji Atarashi, Junichi Nishimura, Tatsuichiro Shima, Yoshinori Umesaki, Masahiro Yamamoto, Masaharu Onoue, Hideo Yagita, Naoto Ishii, Richard Evans, Kenya Honda, and Kiyoshi Takeda. ATP drives lamina propria T(H)17 cell differentiation. *Nature*, 455(7214):808–12, October 2008. ISSN 1476-4687. doi: 10.1038/nature07240. 15

[36] Thomas B Clarke, Kimberly M Davis, Elena S Lysenko, Alice Y Zhou, Yimin Yu, and Jeffrey N Weiser. Recognition of peptidoglycan from the microbiota by Nod1 enhances systemic innate immunity. *Nature Medicine*, 16(2):228–231, February 2010. ISSN 1078-8956. doi: 10.1038/nm.2087. 15, 18

[37] Claudia Di Giacinto, Mariarosaria Marinaro, Massimo Sanchez, Warren Strober, and Monica Boirivant. Probiotics ameliorate recurrent Th1-mediated murine colitis by inducing IL-10 and IL-10-dependent TGF-beta-bearing regulatory cells. *Journal of immunology (Baltimore, Md. : 1950)*, 174(6):3237–46, March 2005. ISSN 0022-1767. 15

[38] Jason A Hall, Nicolas Bouladoux, Cheng Ming Sun, Elizabeth A Wohlfert, Rebecca B Blank, Qing Zhu, Michael E Grigg, Jay A Berzofsky, and Yasmine Belkaid. Commensal DNA limits regulatory T cell conversion and is a natural adjuvant of intestinal immune responses. *Immunity*, 29(4):637–49, October 2008. ISSN 1097-4180. doi: 10.1016/j.immuni.2008.08.009. 15

[39] June L. Round and Sarkis K. Mazmanian. Inducible Foxp3+ regulatory T-cell development by a commensal bacterium of the intestinal microbiota. *Proceedings of the National Academy of Sciences*, 107(27): 12204–12209, July 2010. doi: 10.1073/pnas.0909122107. 16

[40] Valérie Gaboriau-Routhiau, Sabine Rakotobe, Emelyne Lécuyer, Imke Mulder, Annaïg Lan, Chantal Bridonneau, Violaine Rochet, Annamaria Pisi, Marianne De Paepe, Giovanni Brandi, Gérard Eberl, Johannes Snel, Denise Kelly, and Nadine Cerf-Bensussan. The key role of segmented filamentous bacteria in the coordinated maturation of gut helper T cell responses. *Immunity*, 31(4):677–89, October 2009. ISSN 1097-4180. doi: 10.1016/j.immuni.2009.08.020. 16

[41] Ivaylo I Ivanov, Koji Atarashi, Nicolas Manel, Eoin L Brodie, Tatsuichiro Shima, Ulas Karaoz, Dongguang Wei, Katherine C Goldfarb, Clark A Santee, Susan V Lynch, Takeshi Tanoue, Akemi Imaoka, Kikuji Itoh, Kiyoshi Takeda, Yoshinori Umesaki, Kenya Honda, and Dan R Littman. Induction of intestinal Th17 cells by segmented filamentous bacteria. *Cell*, 139(3):485–498, October 2009. doi: 10.1016/j.cell.2009.09.033. 16

[42] Harry Sokol, Bénédicte Pigneur, Laurie Watterlot, Omar Lakhdari, Luis G Bermúdez-Humarán, Jean-Jacques Gratadoux, Sébastien Blugeon, Chantal Bridonneau, Jean-Pierre Furet, Gérard Corthier, Corinne Grangette, Nadia Vasquez, Philippe Pochart, Germain Trugnan, Ginette Thomas, Hervé M Blottière, Joël Doré, Philippe Marteau, Philippe Seksik, and Philippe Langella. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proceedings of the National Academy of Sciences of the United States of America*, 105(43):16731–6, October 2008. ISSN 1091-6490. doi: 10.1073/pnas.0804812105. 16

[43] Giuseppe Mancuso, Angelina Midiri, Carmelo Biondo, Concetta Beninati, Maria Gambuzza, Daniele Macrì, Antonella Bellantoni, Andrej Weintraub, Terje Espevik, and Giuseppe Teti. Bacteroides fragilis-derived lipopolysaccharide produces cell activation and lethal toxicity via toll-like receptor 4. *Infection and immunity*, 73(9):5620–7, September 2005. ISSN 0019-9567. doi: 10.1128/IAI.73.9.5620-5627.2005. 16

[44] Joel V Weinstock, Robert W Summers, and David E Elliott. Role of helminths in regulating mucosal inflammation. *Springer Seminars in Immunopathology*, pages 249–271, 2005. doi: 10.1007/s00281-005-0209-3. 17

[45] G Gibson and M Roberfroid. Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics. *Journal of Nutrition*, January 1995. 17

[46] Daniel C Baumgart and Simon R Carding. Inflammatory bowel disease: cause and immunobiology. *Lancet*, 369(9573):1627–1640, May 2007. doi: 10.1016/S0140-6736(07)60750-8. 17, 24

[47] Bärbel Stecher, Riccardo Robbiani, Alan W Walker, Astrid M Westendorf, Manja Barthel, Marcus Kremer, Samuel Chaffron, Andrew J Macpherson, Jan Buer, Julian Parkhill, Gordon Dougan, Christian von Mering, and Wolf-Dietrich Hardt. Salmonella enterica serovar typhimurium exploits inflammation to compete with the intestinal microbiota. *PLoS biology*, 5(10):2177–89, October 2007. ISSN 1545-7885. doi: 10.1371/journal.pbio.0050244. 17

[48] Sarkis K Mazmanian, Cui Hua Liu, Arthur O Tzianabos, and Dennis L Kasper. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell*, 122(1):107–118, July 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2005.05.007. 18

[49] Jussi Vaahtovuo, Eveliina Munukka, Mika Korkeamäki, Reijo Luukkainen, and Paavo Toivanen. Fecal microbiota in early rheumatoid arthritis. *The Journal of rheumatology*, 35(8):1500–5, August 2008. ISSN 0315-162X. 19

[50] Mikael Kuitunen, Kaarina Kukkonen, Kaisu Juntunen-Backman, Riitta Korpela, Tuija Poussa, Tuula Tuure, Tari Haahtela, and Erkki Savilahti. Probiotics prevent IgE-associated allergy until age 5 years in cesarean-delivered children but not in the total cohort. *The Journal of allergy and clinical immunology*, 123(2):335–41, February 2009. ISSN 1097-6825. doi: 10.1016/j.jaci.2008.11.019. 19

[51] Mohamed Elfatih H. Bashir, Steve Louie, Hai Ning Shi, and Cathryn Nagler-Anderson. Toll-Like Receptor 4 Signaling by Intestinal Microbes Influences Susceptibility to Food Allergy. *J. Immunol.*, 172(11):6978–6987, June 2004. 19

[52] Bengt Björkstén. Diverse microbial exposure - Consequences for vaccine development. *Vaccine*, November 2011. ISSN 1873-2518. doi: 10.1016/j.vaccine.2011.10.074. 19

[53] R Lagos, A Fasano, S S Wasserman, V Prado, O San Martin, P Abrego, G A Losonsky, S Alegria, and M M Levine. Effect of small bowel bacterial overgrowth on the immunogenicity of single-dose live oral cholera vaccine CVD 103-HgR. *The Journal of infectious diseases*, 180(5):1709–12, November 1999. ISSN 0022-1899. doi: 10.1086/315051. 19

[54] Rosana B R Ferreira, L Caetano M Antunes, and B Brett Finlay. Should the human microbiome be considered when developing vaccines? *PLoS pathogens*, 6(11):e1001190, January 2010. ISSN 1553-7374. doi: 10.1371/journal.ppat.1001190. 19

[55] Erick Cardenas and James M Tiedje. New tools for discovering and characterizing microbial diversity. *Current Opinion in Biotechnology*, 19(6):544–549, December 2008. doi: 10.1016/j.copbio.2008.10.010. 21, 72

[56] Silvia G Acinas, Vanja Klepac-Ceraj, Dana E Hunt, Chanathip Pharino, Ivica Ceraj, Daniel L Distel, and Martin F Polz. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, 430(6999):551–554, July 2004. doi: 10.1038/nature02649. 21

[57] A Fernández, Suiying Huang, Sherry Seston, Jian Xing, Robert Hickey, C Criddle, J Tiedje, and A N A Ferna. How stable is stable? Function versus community composition. *Appl Environ Microbiol*, 65(8):3697–3704, August 1999. 21, 72

[58] P Straight and R Kolter. Interspecies Chemical Communication in Bacterial Development. *Annu Rev Microbiol*, June 2009. doi: 10.1146/annurev.micro.091208.073248. 21

[59] Ruth E Ley, Catherine A Lozupone, Micah Hamady, Rob Knight, and Jeffrey I Gordon. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol*, 6(10):776–788, October 2008. doi: 10.1038/nrmicro1978. 21

[60] M J Carter, A J Lobo, and S P L Travis. Guidelines for the management of inflammatory bowel disease in adults. *Gut*, 53 Suppl 5(suppl_5):V1–16, September 2004. ISSN 0017-5749. doi: 10.1136/gut.2004.043372. 24

[61] Michael D Kappelman and Athos Bousvaros. Nutritional concerns in pediatric inflammatory bowel disease patients. *Molecular nutrition & food research*, 52(8):867–74, August 2008. ISSN 1613-4133. doi: 10.1002/mnfr.200700156. 24

[62] R K Yantiss and R D Odze. Diagnostic difficulties in inflammatory bowel disease pathology. *Histopathology*, 48(2):116–32, January 2006. ISSN 0309-0167. doi: 10.1111/j.1365-2559.2005.02248.x. 24

[63] Janice B. Heikenen, Steven L. Werlin, Christopher W. Brown, and Jane P. Balint. Presenting symptoms and diagnostic lag in children with inflammatory bowel disease. *Inflammatory bowel diseases*, 5(3):158–60, August 1999. ISSN 1078-0998. doi: 10.1002/ibd.3780050303. 24

[64] C Spray, G D Debelle, and M S Murphy. Current diagnosis, management and morbidity in paediatric inflammatory bowel disease. *Acta paediatrica (Oslo, Norway : 1992)*, 90(4):400–5, April 2001. ISSN 0803-5253. doi: 10.1111/j.1651-2227.2001.tb00439.x. 24

[65] G J Devroede, W F Taylor, W G Sauer, R J Jackman, and G B Stickler. Cancer risk and life expectancy of children with ulcerative colitis. *The New England journal of medicine*, 285(1):17–21, July 1971. ISSN 0028-4793. doi: 10.1056/NEJM197107012850103. 24

[66] M Peeters, S Joossens, S Vermeire, R Vlietinck, X Bossuyt, and P Rutgeerts. Diagnostic value of anti-Saccharomyces cerevisiae and antineutrophil cytoplasmic autoantibodies in inflammatory bowel disease. *The American journal of gastroenterology*, 96(3):730–4, March 2001. ISSN 0002-9270. doi: 10.1111/j.1572-0241.2001.03613.x. 24

[67] Kjel Andersen, Christoph Vogt, Dirk Blondin, Andreas Beck, Wolfram Heinen, Volker Aurich, Dieter Häussinger, Ulrich Mödder, and Mathias Cohnen. Multi-detector CT-colonography in inflammatory bowel disease: prospective analysis of CT-findings to high-resolution video colonoscopy. *European journal of radiology*, 58(1):140–6, April 2006. ISSN 0720-048X. doi: 10.1016/j.ejrad.2005.11.004. 24

[68] M Löffler, M Weckesser, C Franzius, O Schober, and K-P Zimmer. High diagnostic value of 18F-FDG-PET in pediatric patients with chronic inflammatory bowel disease. *Annals of the New York Academy of Sciences*, 1072:379–85, August 2006. ISSN 0077-8923. doi: 10.1196/annals.1326.014. 24

[69] James D. Lewis. The Utility of Biomarkers in the Diagnosis and Therapy of Inflammatory Bowel Disease. *Gastroenterology*, 140(6):1817–1826.e2, May 2011. ISSN 00165085. doi: 10.1053/j.gastro.2010.11.058. 24, 54

[70] F M Ruemmele, S R Targan, G Levy, M Dubinsky, J Braun, and E G Seidman. Diagnostic accuracy of serological assays in pediatric inflammatory bowel disease. *Gastroenterology*, 115(4):822–9, October 1998. ISSN 0016-5085. 24

[71] Gregory L Austin, Hans H Herfarth, and Robert S Sandler. A critical evaluation of serologic markers for inflammatory bowel disease. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*, 5(5):545–7, May 2007. ISSN 1542-7714. doi: 10.1016/j.cgh.2007.03.006. 24

[72] M C Dubinsky, J J Ofman, M Urman, S R Targan, and E G Seidman. Clinical utility of serodiagnostic testing in suspected pediatric inflammatory bowel disease. *The American journal of gastroenterology*, 96(3):758–65, March 2001. ISSN 0002-9270. doi: 10.1111/j.1572-0241.2001.03618.x. 24, 46

[73] Charles N Bernstein and Fergus Shanahan. Disorders of a modern lifestyle: reconciling the epidemiology of inflammatory bowel diseases. *Gut*, 57(9):1185–91, September 2008. ISSN 1468-3288. doi: 10.1136/gut.2007.122143. 24

[74] Judy H Cho. The genetics and immunopathogenesis of inflammatory bowel disease. *Nature Reviews Immunology*, 8(6):458, June 2008. doi: doi:10.1038/nri2340. 24

[75] Kristen O Arseneau, Hiroshi Tamagawa, Theresa T Pizarro, and Fabio Cominelli. Innate and adaptive immune responses related to IBD pathogenesis. *Current gastroenterology reports*, 9(6):508–512, December 2007. ISSN 1534-312X. 24

[76] Johan Dicksved, Jonas Halfvarson, Magnus Rosenquist, Gunnar Järnerot, Curt Tysk, Juha Apajalahti, Lars Engstrand, and Janet K Jansson. Molecular analysis of the gut microbiota of identical twins with Crohn's disease. *The ISME journal*, 2(7):716–27, July 2008. ISSN 1751-7370. doi: 10.1038/ismej.2008.37. 24, 25, 52

[77] R Balfour Sartor. Microbial influences in inflammatory bowel diseases. *Gastroenterology*, 134(2):577–94, February 2008. ISSN 1528-0012. doi: 10.1053/j.gastro.2007.11.059. 24, 38, 53

[78] Sandra C Kim, Susan L Tonkonogy, Carol A Albright, Julia Tsang, Edward J Balish, Jonathon Braun, Mark M Huycke, and R Balfour Sartor. Variable phenotypes of enterocolitis in interleukin 10-deficient mice monoassociated with two different commensal bacteria. *Gastroenterology*, 128(4):891–906, May 2005. ISSN 0016-5085. 24

[79] Warwick Selby, Paul Pavli, Brendan Crotty, Tim Florin, Graham Radford-Smith, Peter Gibson, Brent Mitchell, William Connell, Robert Read, Michael Merrett, Hooi Ee, and David Hetzel. Two-year combination antibiotic therapy with clarithromycin, rifabutin, and clofazimine for Crohn's disease. *Gastroenterology*, 132(7):2313–9, June 2007. ISSN 0016-5085. doi: 10.1053/j.gastro.2007.03.031. 25

[80] S K Böhm and W Kruis. Probiotics: do they help to control intestinal inflammation? *Annals of the New York Academy of Sciences*, 1072:339–50, August 2006. ISSN 0077-8923. doi: 10.1196/annals.1326.005. 25

[81] Ben P Willing, Johan Dicksved, Jonas Halfvarson, Anders F Andersson, Marianna Lucio, Zongli Zheng, Gunnar Järnerot, Curt Tysk, Janet K Jansson, and Lars Engstrand. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology*, 139(6):1844–1854.e1, December 2010. ISSN 1528-0012. doi: 10.1053/j.gastro.2010.08.049. 25, 53

[82] Pierre Baldi and Sø ren Brunak. *Bioinformatics: the machine learning approach*. August 2001. ISBN 0-262-02506-X. 26

[83] Margaret A Shipp, Ken N Ross, Pablo Tamayo, Andrew P Weng, Jeffery L Kutok, Ricardo C T Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S Pinkus, Tane S Ray, Margaret A Koval, Kim W Last, Andrew Norton, T Andrew Lister, Jill Mesirov, Donna S Neuberg, Eric S Lander, Jon C Aster, and Todd R Golub. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, 8(1):68–74, January 2002. ISSN 1078-8956. doi: 10.1038/nm0102-68. 26

[84] Dan Knights, Elizabeth K Costello, and Rob Knight. Supervised classification of human microbiota. *FEMS microbiology reviews*, pages 1–17, September 2010. ISSN 1574-6976. doi: 10.1111/j.1574-6976.2010.00251.x. 26, 52, 68, 69, 75, 91

[85] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002. 26, 65, 75, 93

[86] A Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997. ISSN 00313203. doi: 10.1016/S0031-3203(96)00142-2. 26

[87] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–5, August 2003. ISSN 0027-8424. doi: 10.1073/pnas.1530509100. 31, 41, 42, 49, 66

[88] Dan Turner, Jeffrey Hyams, James Markowitz, Trudy Lerer, David R Mack, Jonathan Evans, Marian Pfefferkorn, Joel Rosh, Marsha Kay, Wallace Crandall, David Keljo, Anthony R Otley, Subra Kugathasan, Ryan Carvalho, Maria Oliva-Hemker, Christine Langton, Petar Mamula, Athos Bousvaros, Neal LeLeiko, and Anne M Griffiths. Appraisal of the pediatric ulcerative colitis activity index (PUCAI). *Inflammatory bowel diseases*, 15(8):1218–23, August 2009. ISSN 1536-4844. doi: 10.1002/ibd.20867. 34, 35, 56

[89] Dan Turner, Anne M Griffiths, Thomas D Walters, Tong Seah, James Markowitz, Marian Pfefferkorn, David Keljo, Anthony Otley, Neal S Leleiko, David Mack, Jeffrey Hyams, and Arie Levine. Appraisal of the pediatric Crohn's disease activity index on four prospectively collected datasets: recommended cutoff values and clinimetric properties. *The American journal of gastroenterology*, 105(9):2085–92, September 2010. ISSN 1572-0241. doi: 10.1038/ajg.2010.143. 34, 35, 56

[90] Anne M Griffiths, Anthony R Otley, Jeffrey Hyams, Antonio R Quiros, Richard J Grand, Athos Bousvaros, Brian G Feagan, and George R Ferry. A review of activity indices and end points for clinical trials in children with Crohn's disease. *Inflammatory bowel diseases*, 11(2):185–96, February 2005. ISSN 1078-0998. 34

[91] C P Tamboli, C Neut, P Desreumaux, and J F Colombel. Dysbiosis in inflammatory bowel disease. *Gut*, 53 (1):1–4, January 2004. ISSN 0017-5749. 38, 53

[92] Shadi Sepehri, Roman Kotlowski, Charles N Bernstein, and Denis O Krause. Microbial diversity of inflamed and noninflamed gut biopsy tissues in inflammatory bowel disease. *Inflammatory bowel diseases*, 13(6): 675–83, June 2007. ISSN 1078-0998. doi: 10.1002/ibd.20101. 41, 52

[93] Patrick D Schloss. Evaluating different approaches that test whether microbial communities have the same structure. *The ISME Journal*, 2(3):265, January 2008. doi: doi:10.1038/ismej.2008.5. 63

[94] Brian J Haas, Dirk Gevers, Ashlee Earl, Mike Feldgarden, Doyle V Ward, Georgia Giannokous, Dawn Ciulla, Diana Tabbaa, Sarah K Highlander, Erica Sodergren, Barbara Methe, Todd Z Desantis, Joseph F Petrosino, Rob Knight, and Bruce W Birren. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome research*, 21(3):494–504, January 2011. ISSN 1549-5469. doi: 10.1101/gr.112730.110. 63

[95] J R Cole, Q Wang, E Cardenas, J Fish, B Chai, R J Farris, A S Kulam-Syed-Mohideen, D M McGarrell, T Marsh, G M Garrity, and J M Tiedje. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic acids research*, 37(Database issue):D141–5, January 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn879. 65

[96] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73 (16):5261–7, August 2007. ISSN 0099-2240. doi: 10.1128/AEM.00062-07. 65

[97] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.r-project.org`. 65, 66

[98] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN 978-0-387-98140-6. 65, 66

[99] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)*, 23(19):2507–17, October 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm344. 66

[100] Cara M Santelli, Beth N Orcutt, Erin Banning, Wolfgang Bach, Craig L Moyer, Mitchell L Sogin, Hubert Staudigel, and Katrina J Edwards. Abundance and diversity of microbial life in ocean crust. *Nature*, 453 (7195):653–6, May 2008. ISSN 1476-4687. doi: 10.1038/nature06899. 69

[101] Silvia A Piñeiro, Henry N Williams, and O Colin Stine. Phylogenetic relationships amongst the saltwater members of the genus Bacteriovorax using rpoB sequences and reclassification of Bacteriovorax stolpii as Bacteriolyticum stolpii gen. nov., comb. nov. *International journal of systematic and evolutionary microbiology*, 58(Pt 5):1203–9, May 2008. ISSN 1466-5026. doi: 10.1099/ijs.0.65710-0. 69

[102] Olivia U Mason, Carol A Di Meo-Savoie, Joy D Van Nostrand, Jizhong Zhou, Martin R Fisk, and Stephen J Giovannoni. Prokaryotic diversity, distribution, and insights into their role in biogeochemical cycling in marine basalts. *The ISME journal*, 3(2):231–42, February 2009. ISSN 1751-7370. doi: 10.1038/ismej.2008.92. 69, 71

[103] Cara M Santelli, Virginia P Edgcomb, Wolfgang Bach, and Katrina J Edwards. The diversity and abundance of bacteria inhabiting seafloor lavas positively correlate with rock alteration. *Environmental microbiology*, 11 (1):86–98, January 2009. ISSN 1462-2920. doi: 10.1111/j.1462-2920.2008.01743.x. 69

[104] Jennifer L Bowen, Byron C Crump, Linda A Deegan, and John E Hobbie. Salt marsh sediment bacteria: their distribution and response to external nutrient inputs. *The ISME journal*, 3(8):924–34, August 2009. ISSN 1751-7370. doi: 10.1038/ismej.2009.44. 69

[105] Stephanie A Smith, James N Benardini, Janice L Strap, and Ronald L Crawford. Diversity of aerobic and facultative alkalitolerant and halotolerant endospore formers in soil from the Alvord Basin, Oregon. *Systematic and applied microbiology*, 32(4):233–44, July 2009. ISSN 1618-0984. doi: 10.1016/j.syapm.2008.09.008. 69

[106] Tara A Gianoulis, Jeroen Raes, Prianka V Patel, Robert Bjornson, Jan O Korbel, Ivica Letunic, Takuji Yamada, Alberto Paccanaro, Lars J Jensen, Michael Snyder, Peer Bork, and Mark B Gerstein. Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci USA*, 106(5):1374–1379, February 2009. ISSN 1091-6490. doi: 10.1073/pnas.0808022106. 69

[107] Anat Kreimer, Elhanan Borenstein, Uri Gophna, and Eytan Ruppin. The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences*, 105(19):6976, May 2008. doi: 10.1073/pnas.0712149105. 69

[108] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804–810, October 2007. doi: 10.1038/nature06244. 69, 72

[109] Masahira Hattori and Todd D Taylor. The human intestinal microbiome: a new frontier of human biology. *DNA Res*, 16(1):1–12, February 2009. doi: 10.1093/dnares/dsn033. 69, 72

[110] Mihai Pop. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, 10(4): 354–66, July 2009. ISSN 1477-4054. doi: 10.1093/bib/bbp026. 69, 72

[111] Victor Kunin, Alex Copeland, Alla Lapidus, Konstantinos Mavromatis, and Philip Hugenholtz. A Bioinformatician's Guide to Metagenomics. *Microbiol. Mol. Biol. Rev.*, 72(4):557–578, December 2008. ISSN 1098-5557. doi: 10.1128/MMBR.00009-08. 69

[112] W G Weisburg, S M Barns, D A Pelletier, and D J Lane. 16S ribosomal DNA amplification for phylogenetic study. *J Bacteriol*, 173(2):697–703, January 1991. 70

[113] Rebecca J Case, Yan Boucher, Ingela Dahllöf, Carola Holmström, W Ford Doolittle, and Staffan Kjelleberg. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol*, 73(1):278–288, January 2007. doi: 10.1128/AEM.01177-06. 71

[114] Susannah G Tringe and Philip Hugenholtz. A renaissance for the pioneering 16S rRNA gene. *Current Opinion in Microbiology*, 11(5):442–446, October 2008. doi: 10.1016/j.mib.2008.09.011. 71, 72

[115] William P Hanage, Christophe Fraser, and Brian G Spratt. Sequences, sequence clusters and bacterial species. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, 361(1475): 1917–27, 2006. ISSN 09628436. doi: 10.1098/rstb.2006.1917. 71

[116] Cynthia J Bishop, David M Aanensen, Gregory E Jordan, Mogens Kilian, William P Hanage, and Brian G Spratt. Assigning strains to bacterial species via the internet. *BMC Biology*, 7(1):3, 2009. 71

[117] Daniel MacLean, Jonathan D. G. Jones, and David J. Studholme. Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Reviews Microbiology*, 7(4):287–296, February 2009. ISSN 1740-1526. doi: 10.1038/nrmicro2088. 72

[118] Peter J. Russell. *iGenetics: A Molecular Approach*. Pearson, 2005. ISBN 0321312074. 72

[119] Catherine A Lozupone and Rob Knight. Species divergence and the measurement of microbial diversity. *FEMS Microbiol Rev*, 32(4):557–578, July 2008. doi: 10.1111/j.1574-6976.2008.00111.x. 72

[120] Andrew P Martin. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied and environmental microbiology*, 68(8):3673–82, August 2002. ISSN 0099-2240. 73

[121] D E Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989. ISBN 0201157675. 74

[122] Elizabeth K. Costello, Christian L. Lauber, Micah Hamady, Noah Fierer, Jeffrey I. Gordon, and Rob Knight. Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science*, 326(5960): 1694–1697, December 2009. ISSN 1095-9203. doi: 10.1126/science.1177486. 75

[123] Chiachi Hwang, Weimin Wu, Terry J Gentry, Jack Carley, Gail A Corbin, Sue L Carroll, David B Watson, Phil M Jardine, Jizhong Zhou, Craig S Criddle, and Matthew W Fields. Bacterial community succession during in situ uranium bioremediation: spatial similarities along controlled flow paths. *The ISME journal*, 3 (1):47–64, January 2009. ISSN 1751-7370. doi: 10.1038/ismej.2008.77. 75

[124] Chao Chen and Andy Liaw. Using random forest to learn imbalanced data. *Discovery*, (1999):1–12, 2004. 85, 93

[125] Catherine a Lozupone, Micah Hamady, Scott T Kelley, and Rob Knight. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and environmental microbiology*, 73(5):1576–85, March 2007. ISSN 0099-2240. doi: 10.1128/AEM.01996-06. 86

[126] Susannah Green Tringe, Christian von Mering, Arthur Kobayashi, Asaf A Salamov, Kevin Chen, Hwai W Chang, Mircea Podar, Jay M Short, Eric J Mathur, John C Detter, Peer Bork, Philip Hugenholtz, and Edward M Rubin. Comparative metagenomics of microbial communities. *Science (New York, N.Y.)*, 308(5721): 554–7, April 2005. ISSN 1095-9203. doi: 10.1126/science.1107851. 86

[127] Catherine Lozupone, Micah Hamady, and Rob Knight. UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, 7:371, January 2006. doi: 10.1186/1471-2105-7-371. 87

[128] Alexandros Stamatakis, Michael Ott, Thomas Ludwig, and Victor Malyshkin. *Parallel Computing Technologies*, volume 3606 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-28126-9. doi: 10.1007/11535294. 90

[129] Dan Knights, Justin Kuczynski, Omry Koren, Ruth E Ley, Dawn Field, Rob Knight, Todd Z DeSantis, and Scott T Kelley. Supervised classification of microbiota mitigates mislabeling errors. *The ISME journal*, 5(4):570–3, April 2011. ISSN 1751-7370. doi: 10.1038/ismej.2010.148. 91

[130] Dan Knights, Justin Kuczynski, Emily S Charlson, Jesse Zaneveld, Michael C Mozer, Ronald G Collman, Frederic D Bushman, Rob Knight, and Scott T Kelley. Bayesian community-wide culture-independent microbial source tracking. *Nature methods*, 8(9):761–3, September 2011. ISSN 1548-7105. doi: 10.1038/nmeth.1650. 91

[131] Zhenqiu Liu, William Hsiao, Brandi L Cantarel, Elliott Franco Drábek, and Claire Fraser-Liggett. Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics (Oxford, England)*, 27(23):3242–9, December 2011. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr547. 91

[132] Jennifer Carpenter. May the best analyst win. *Science (New York, N.Y.)*, 331(6018):698–9, February 2011. ISSN 1095-9203. doi: 10.1126/science.331.6018.698. 91, 162

[133] V N Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999. 92

[134] Vladimir Naumovich Vapnik. *The nature of statistical learning theory*, volume 8 of *Statistics for engineering and information science*. Springer, 2000. ISBN 0387987800. doi: 10.1109/TNN.1997.641482. 92

[135] T G Dietterich. Ensemble methods in machine learning. In *First International Workshop on Multiple Classifier Systems*, volume 1857 of *LNCS*, pages 1–15. Springer, 2000. URL `http://www.springerlink.com/index/lg5742lr8810612k.pdf`. 93

[136] Thomas G Dietterich. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees : Bagging , Boosting , and Randomization. *Machine Learning*, 40(2):139–157, 2000. ISSN 08856125. doi: 10.1023/A:1007607513941. 93

[137] Leo Breiman. Random Forest. *Machine learning*, pages 1–33, 2001. 93

[138] Susan K Pierce. Lipid rafts and B-cell activation. *Nature Reviews Immunology*, 2(2):96–105, February 2002. doi: 10.1038/nri726. 98, 111, 114, 116

[139] Pavel Tolar, Hae Won Sohn, and Susan K Pierce. Viewing the antigen-induced initiation of B-cell activation in living cells. *Immunol Rev*, 221:64–76, February 2008. doi: 10.1111/j.1600-065X.2008.00583.x. 98

[140] Marcus R Clark, Don Massenburg, Miao Zhang, and Karyn Siemasko. Molecular mechanisms of B cell antigen receptor trafficking. *Ann N Y Acad Sci*, 987:26–37, April 2003. 98

[141] Marcus R Clark, Donald Massenburg, Karyn Siemasko, Ping Hou, and Miao Zhang. B-cell antigen receptor signaling requirements for targeting antigen to the MHC class II presentation pathway. *Curr Opin Immunol*, 16(3):382–387, June 2004. doi: 10.1016/j.coi.2004.03.007. 98

[142] You-Me Kim, Jennifer Yi-Jiun Pan, Gregory A Korbel, Victor Peperzak, Marianne Boes, and Hidde L Ploegh. Monovalent ligation of the B cell receptor induces receptor activation but fails to promote antigen presentation. *Proc Natl Acad Sci USA*, 103(9):3327–3332, February 2006. doi: 10.1073/pnas.0511315103. 98, 105, 111, 114

[143] C Langlet, A M Bernard, P Drevot, and H T He. Membrane rafts and signaling by the multichain immune recognition receptors. *Curr Opin Immunol*, 12(3):250–255, June 2000. 98

[144] K Simons and D Toomre. Lipid rafts and signal transduction. *Nat Rev Mol Cell Biol*, 1(1):31–39, October 2000. doi: 10.1038/35036052. 98

[145] Michelle Dykstra, Anu Cherukuri, Hae Won Sohn, Shiang-Jong Tzeng, and Susan K Pierce. Location is everything: lipid rafts and immune cell signaling. *Annual Review of Immunology*, 21(3):457–481, January 2003. ISSN 0732-0582. doi: 10.1146/annurev.immunol.21.120601.141021. 98

[146] M Reth and J Wienands. Initiation and processing of signals from the B cell antigen receptor. *Annual Review of Immunology*, 15:453–479, January 1997. doi: 10.1146/annurev.immunol.15.1.453. 98

[147] Ping Hou, Elizabeth Araujo, Tong Zhao, Miao Zhang, Don Massenburg, Margaret Veselits, Colleen Doyle, Aaron R. Dinner, and Marcus R. Clark. B Cell Antigen Receptor Signaling and Internalization Are Mutually Exclusive Events. *PLoS Biology*, 4(7), July 2006. doi: 10.1371/journal.pbio.0040200. 98, 105, 117

[148] Mathieu Bléry, Lina Tze, Lisa A Miosge, Jesse E Jun, and Christopher C Goodnow. Essential role of membrane cholesterol in accelerated BCR internalization and uncoupling from NF-kappa B in B cell clonal anergy. *J Exp Med*, 203(7):1773–1783, July 2006. doi: 10.1084/jem.20060552. 98, 105, 116, 117

[149] Hae Won Sohn, Pavel Tolar, Tian Jin, and Susan K Pierce. Fluorescence resonance energy transfer in living cells reveals dynamic membrane changes in the initiation of B cell signaling. *Proc Natl Acad Sci USA*, 103 (21):8143–8148, May 2006. doi: 10.1073/pnas.0509858103. 98, 160

[150] Pavel Tolar, Hae Won Sohn, and Susan K Pierce. The initiation of antigen-induced B cell antigen receptor signaling viewed in living cells by fluorescence resonance energy transfer. *Nature Immunology*, 6(11): 1168–1176, November 2005. doi: 10.1038/ni1262. 98, 116, 160

[151] Raibatak Das, Stephanie Hammond, David Holowka, and Barbara Baird. Real-time cross-correlation image analysis of early events in IgE receptor signaling. *Biophysical Journal*, page biophysj.107.105502v1, March 2008. doi: 10.1529/biophysj.107.105502. 107

[152] Neetu Gupta and Anthony L DeFranco. Visualizing lipid raft dynamics and early signaling events during antigen receptor-mediated B-lymphocyte activation. *Mol Biol Cell*, 14(2):432–444, February 2003. doi: 10.1091/mbc.02-05-0078. 114

[153] W W Schamel and M Reth. Monomeric and oligomeric complexes of the B cell antigen receptor. *Immunity*, 13(1):5–14, July 2000. 116

[154] G M Whitesides, E Ostuni, S Takayama, X Jiang, and D E Ingber. Soft lithography in biology and biochemistry. *Annual review of biomedical engineering*, 3:335–373, January 2001. doi: 10.1146/annurev.bioeng.3.1.335. 118

[155] Jamil El-Ali, Peter K Sorger, and Klavs F Jensen. Cells on chips. *Nature*, 442(7101):403–411, July 2006. doi: 10.1038/nature05063. 119, 161

[156] A Atherton and G V Born. Quantitative investigations of the adhesiveness of circulating polymorphonuclear leucocytes to blood vessel walls. *J Physiol (Lond)*, 222(2):447–474, April 1972. 121

[157] Yolanda R Carrasco and Facundo D Batista. B-cell activation by membrane-bound antigens is facilitated by the interaction of VLA-4 with VCAM-1. *EMBO J*, 25(4):889–899, February 2006. doi: 10.1038/sj.emboj.7600944. 121

[158] R M Zinkernagel. Immunology taught by viruses. *Science (New York, N.Y.)*, 271(5246):173–8, January 1996. ISSN 0036-8075. 129

[159] Dennis R Burton. Antibodies, viruses and vaccines. *Nature reviews. Immunology*, 2(9):706–13, September 2002. ISSN 1474-1733. doi: 10.1038/nri891. 129

[160] N Jerne and A Nordin. Plaque Formation in Agar by Single Antibody-Producing Cells. *Science*, 140(3565):405, April 1963. doi: 10.1126/science.140.3565.405. 129

[161] N R Klinman and G Aschinazi. The stimulation of splenic foci in vitro. *J Immunol*, 106(5):1338–1344, May 1971. 129

[162] G Kelsoe and J T Stout. Cloning of mitogen- and antigen-reactive B lymphocytes on filter paper discs. II. Paratope frequencies within the mitogen-selected repertoire. *Cell Immunol*, 98(2):506–516, April 1986. 129

[163] G Kelsoe. Cloning of mitogen- and antigen-reactive B lymphocytes on filter paper discs. I. A description of the technique and of methods for the analysis of colonies. *J Immunol Methods*, 76(2):345–363, February 1985. 129

[164] C C Czerkinsky, L A Nilsson, H Nygren, O Ouchterlony, and A Tarkowski. A solid-phase enzyme-linked immunospot (ELISPOT) assay for enumeration of specific antibody-secreting cells. *J Immunol Methods*, 65(1-2):109–121, December 1983. 129

[165] K Hayakawa, R Ishii, K Yamasaki, T Kishimoto, and R R Hardy. Isolation of high-affinity memory B cells: phycoerythrin as a probe for antigen-binding cells. *Proc Natl Acad Sci USA*, 84(5):1379–1383, March 1987. 129, 150

[166] M G McHeyzer-Williams, M J McLean, P A Lalor, and G J Nossal. Antigen-driven B cell differentiation in vivo. *The Journal of experimental medicine*, 178(1):295–307, July 1993. ISSN 0022-1007. 129, 150

[167] J Christopher Love, Jehnna L Ronan, Gijsbert M Grotenbreg, Annemarthe G van der Veen, and Hidde L Ploegh. A microengraving method for rapid selection of single cells producing antigen-specific antibodies. *Nat Biotechnol*, 24(6):703–707, June 2006. doi: 10.1038/nbt1210. 130, 131, 152, 153

[168] Victor C Rucker, Karen L Havenstrite, and Amy E Herr. Antibody microarrays for native toxin detection. *Anal Biochem*, 339(2):262–270, April 2005. doi: 10.1016/j.ab.2005.01.030. 131

[169] Jefferson Foote, Herman N Eisent, and H N Eisen. Kinetic and affinity limits on antibodies produced during immune responses. *Proc Natl Acad Sci USA*, 92(5):1254–1256, February 1995. 149

[170] Rino Rappuoli. Bridging the knowledge gaps in vaccine design. *Nat Biotechnol*, 25(12):1361–1366, December 2007. doi: 10.1038/nbt1207-1361. 150

[171] Timothy J Bradford, Xiaoju Wang, and Arul M Chinnaiyan. Cancer immunomics: using autoantibody signatures in the early detection of prostate cancer. *Urol Oncol*, 24(3):237–242, January 2006. doi: 10.1016/j.urolonc.2005.11.033. 151

[172] Ulisses M Braga-Neto, E Marques Jr, Immunomics New Challenges, Old Problems, and Ernesto T A Marques. From Functional Genomics to Functional Immunomics: New Challenges, Old Problems, Big Rewards. *PLoS Comput Biol*, 2(7):e81, July 2006. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0020081. 151

[173] Michel Caron, Geneviève Choquet-Kastylevsky, and Raymonde Joubert-Caron. Cancer immunomics using autoantibody signatures for biomarker discovery. *Mol Cell Proteomics*, 6(7):1115–1122, July 2007. ISSN 1535-9476. doi: 10.1074/mcp.R600016-MCP200. 151

[174] I Cohen. Real and artificial immune systems: computing the state of the body. *Nat Rev Immunol*, January 2007. 151

[175] Julie Hardouin, Jean-Paul Lasserre, Loïk Sylvius, Raymonde Joubert-Caron, and Michel Caron. Cancer immunomics: from serological proteome analysis to multiple affinity protein profiling. *Ann N Y Acad Sci*, 1107:223–230, June 2007. doi: 10.1196/annals.1381.024. 151

[176] Francisco J Quintana, Gad Getz, Guy Hed, Eytan Domany, and Irun R Cohen. Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: a bio-informatic approach to immune complexity. *J Autoimmun*, 21(1):65–75, August 2003. 151

[177] Francisco J Quintana, Peter H Hagedorn, Gad Elizur, Yifat Merbl, Eytan Domany, and Irun R Cohen. Functional immunomics: microarray analysis of IgG autoantibody repertoires predicts the future response of mice to induced diabetes. *Proc Natl Acad Sci USA*, 101 Suppl:14615–14621, October 2004. doi: 10.1073/pnas.0404848101. 151

[178] Sai T Reddy and George Georgiou. Systems analysis of adaptive immunity by utilization of high-throughput technologies. *Current opinion in biotechnology*, 22(4):584–9, August 2011. ISSN 1879-0429. doi: 10.1016/j.copbio.2011.04.015. 158

[179] Jehnna L Ronan, Craig M Story, Eliseo Papa, and J Christopher Love. Optimization of the surfaces used to capture antibodies from single hybridomas reduces the time required for microengraving. *Journal of Immunological Methods*, 340(2):164–169, January 2009. doi: 10.1016/j.jim.2008.10.018. 158

[180] Adebola O Ogunniyi, Craig M Story, Eliseo Papa, Eduardo Guillen, and J Christopher Love. Screening individual hybridomas by microengraving to discover monoclonal antibodies. *Nature protocols*, 4(5):767–82, January 2009. ISSN 1750-2799. doi: 10.1038/nprot.2009.40. 158, 159

[181] Navin Varadarajan, Douglas S Kwon, Kenneth M Law, Adebola O Ogunniyi, Melis N Anahtar, James M Richter, Bruce D Walker, and J Christopher Love. Rapid, efficient functional characterization and recovery of HIV-specific human CD8+ T cells using microengraving. *Proceedings of the National Academy of Sciences of the United States of America*, 109(10):3885–90, February 2012. ISSN 1091-6490. doi: 10.1073/pnas.1111205109. 158, 160

[182] Qing Han, Neda Bagheri, Elizabeth M Bradshaw, David A Hafler, Douglas A Lauffenburger, and J Christopher Love. Polyfunctional responses by human T cells result from sequential release of cytokines. *Proceedings of the National Academy of Sciences of the United States of America*, 109(5):1607–12, January 2012. ISSN 1091-6490. doi: 10.1073/pnas.1117194109. 158

[183] Qing Han, Elizabeth M Bradshaw, Björn Nilsson, David A Hafler, and J Christopher Love. Multidimensional analysis of the frequencies and rates of cytokine secretion from single cells by quantitative microengraving. *Lab on a chip*, 10(11):1391–400, June 2010. ISSN 1473-0197. doi: 10.1039/b926849a. 158, 159

[184] Navin Varadarajan, Boris Julg, Yvonne J Yamanaka, Huabiao Chen, Adebola O Ogunniyi, Elizabeth McAndrew, Lindsay C Porter, Alicja Piechocka-Trocha, Brenna J Hill, Daniel C Douek, Florencia Pereyra, Bruce D Walker, and J Christopher Love. A high-throughput single-cell analysis of human CD8 T cell functions reveals discordance for cytokine secretion and cytolysis. *The Journal of clinical investigation*, 121(11):4322–31, November 2011. ISSN 1558-8238. doi: 10.1172/JCI58653. 159

[185] Elizabeth M Bradshaw, Sally C Kent, Vinay Tripuraneni, Tihamer Orban, Hidde L Ploegh, David A Hafler, and J Christopher Love. Concurrent detection of secreted products from human lymphocytes by microengraving: cytokines and antigen-reactive antibodies. *Clinical immunology (Orlando, Fla.)*, 129(1):10–8, October 2008. ISSN 1521-7035. doi: 10.1016/j.clim.2008.06.009. 160

[186] Jiang F Zhong, Yan Chen, Joshua S Marcus, Axel Scherer, Stephen R Quake, Clive R Taylor, and Leslie P Weiner. A microfluidic processor for gene expression profiling of single human embryonic stem cells. *Lab on a chip*, 8(1):68–74, January 2008. ISSN 1473-0197. doi: 10.1039/b712116d. 160

[187] Doron Gerber, Sebastian J Maerkl, and Stephen R Quake. An in vitro microfluidic approach to generating protein-interaction networks. *Nature methods*, 6(1):71–4, January 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1289. 160

[188] Yael S Schiffenbauer, Yael Kalma, Ella Trubniykov, Orit Gal-Garber, Lilach Weisz, Asaf Halamish, Michael Sister, and Gideon Berke. A cell chip for sequential imaging of individual non-adherent live cells reveals transients and oscillations. *Lab on a chip*, 9(20):2965–72, October 2009. ISSN 1473-0197. doi: 10.1039/b904778f. 160

[189] Benjamin Potsaid, Yves Bellouard, and John Wen. Adaptive Scanning Optical Microscope (ASOM): A multidisciplinary optical microscope design for large field of view and high resolution imaging. *Optics express*, 13(17):6504–18, August 2005. ISSN 1094-4087. 160

[190] Sanjee Abeytunge, Yongbiao Li, Bjorg Larson, Ricardo Toledo-Crow, and Milind Rajadhyaksha. Rapid confocal imaging of large areas of excised tissue with strip mosaicing. *Journal of biomedical optics*, 16(5):050504, May 2011. ISSN 1560-2281. doi: 10.1117/1.3582335. 160

[191] R H Webb and F Rogomentich. Confocal microscope with large field and working distance. *Applied optics*, 38(22):4870–5, August 1999. ISSN 0003-6935. 160

[192] Ayça Yalçın, Yvonne J Yamanaka, and J Christopher Love. Analytical technologies for integrated single-cell analysis of human immune responses. *Methods in molecular biology (Clifton, N.J.)*, 853:211–35, January 2012. ISSN 1940-6029. doi: 10.1007/978-1-61779-567-1_16. 160

[193] Francis Lin and Eugene C Butcher. T cell chemotaxis in a simple microfluidic device. *Lab on a chip*, 6(11):1462–9, November 2006. ISSN 1473-0197. doi: 10.1039/b607071j. 161

[194] Pavel Tolar. Inside the microcluster: antigen receptor signalling viewed with molecular imaging tools. *Immunology*, 133(3):271–7, July 2011. ISSN 1365-2567. doi: 10.1111/j.1365-2567.2011.03452.x. 161

[195] Thomas G. Dietterich and Thomas G Dietterich. Machine-Learning Research – Four Current Directions. *AI MAGAZINE*, 18:97 – 136, 1997. 161

[196] James I Prosser, Brendan J M Bohannan, Tom P Curtis, Richard J Ellis, Mary K Firestone, Rob P Freckleton, Jessica L Green, Laura E Green, Ken Killham, Jack J Lennon, A Mark Osborn, Martin Solan, Christopher J van der Gast, and J Peter W Young. The role of ecological theory in microbial ecology. *Nature reviews. Microbiology*, 5(5):384–92, May 2007. ISSN 1740-1534. doi: 10.1038/nrmicro1643. 163