UC San Diego
UC San Diego Electronic Theses and Dissertations

Title
The human gut microbiome : characterization challenges and HIV disease association

Permalink
https://escholarship.org/uc/item/47j9b3v9

Authors
Pérez Santiago, Josué
Pérez Santiago, Josué

Publication Date
2012

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

The Human Gut Microbiome: Characterization Challenges and HIV Disease Association

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor in Philosophy

in

Bioinformatics and System Biology

by

Josué Pérez Santiago

Committee in charge:

Professor Douglas D. Richman, Chair
Professor Kun Zhang, Co-Chair
Professor Sergei L. Kosakovsky Pond
Professor David M. Smith
Professor Christopher H. Woelk

2012

`

Copyright

Josué Pérez Santiago, 2012

All rights reserved

`

The Dissertation of Josué Pérez Santiago is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____
Co-Chair

____
Chair

University of California, San Diego

2012

DEDICATION

Thinking about one person to dedicate my thesis and all my work throughout these years, the only person that absolutely deserves that recognition is my mom, mi mamá hermosa! Mami, por fin he logrado uno de mis más grandes sueños en la vida y en parte te lo debo a ti. Tu has estado ahí siempre, apoyando y respetando todas la decisiones que he tomado en la vida. Me has dado principalmente tu amor pero más que todo te has sacrificado para que yo haya podido llegar a donde he llegado. Realmente no tengo palabras suficientes para decirte lo agradecido que estoy por como has sido y todo lo que has hecho hasta el día de hoy. Te amo con todo mi corazón y te dedico mi tesis y mi doctorado por ser la mejor mamá hermosa del mundo!!!!

Leaving Puerto Rico was one of the hardest decisions I had to make. It took me long time to get used to the all changes. One person that made the difference in my life here is my best friend Nisha. She is one of the best things that have ever happened to me and because of her I started to feel that I had a family here as well. Nisha, you and I have shared most of our biggest moments of our lives together and I really hope that no matter where we go, we keep sharing those moments forever. You are an awesome friend, and even though every once and then I tell you these things, you know that you are really important in my life and that I love you a bunch! Next year, we will walk together in the graduation!!!

As it was difficult for me to leave Puerto Rico, I think for my family it was even worse… they had to let the baby leave the nest. It was the first time that one of us left the country and not only that, I was going to be completely by myself. Very fortunate for me, my big-little sister Johanna was there and came with me when I moved to make sure that her “little baby” was going to be OK. Johy, realmente yo no se que hubiese hecho si tu no hubieras venido conmigo. La realidad del caso era que yo no tenía idea de a lo que me enfrentaba y si no hubiese sido por ti quizá yo no hubiese tenido ni una cama donde dormir. Siempre has estado ahí, ayudándome, cuidándome y queriendome mucho y eso

es algo que siempre voy a tener conmigo esté donde esté. Me da mucha alegría que así como yo cumplo unos de mis más grandes sueños, tu también estás realizando uno de los tuyos y a la misma vez!!!! Te quiero mucho!! And of course, my sister was not the only who took care of me when I left, all the rest of my family supported in all aspects possible, my dad, my brother and his wife Albert y Lizzie, my sister and her husband Carmencita and Rafa, my new and almost brother Tommy, my best friend Wanda and especially my beautiful nieces, Desirée, Nicole, Pamela and Karina! It has been very difficult for me to see my baby girls grow without me. Desi was 13, Nicole was 12, Pamela was 10 and Karina was not even born at the time I left but somehow our bonds have become stronger these years. Los amo a todos con todo mi corazón y estoy muy agradecido por todo lo cada uno de ustedes ha hecho por mi y por quererme como lo hacen!!!

My six years here have also given me really good friends that thanks to them my life turned out to be really great here. We have made parties, trips, cooking clubs, dancing clubs, Puerto Rican mafia… what can I say… we are very creative… When you are away from home but you find people like Rubén and Bárbara, you don't really feel that away. Everyone should see one of our Thanksgiving dinners so people can understand! Also, my other closest friends here have given me moral support most of the times I had to do some kind of experiment during the weekends, especially those long 454s or helped me when I was writing my thesis… so Gaby, Nancy, Sowbi and Demetri, thanks for your time, support and love!! All of you represent a piece of who I am and I am happy to have all of you. But it wouldn't be fair if I didn't mention one person that has been really important as a friend and to my professional growth, Marta. It was with her that I felt for the first time how important my work was a Bioinformatician and that I was good teaching and explaining things to someone outside my field. She has helped me in so many ways with my thesis that it will take me more than one page to acknowledge her. In her thesis, Marta said that she learned two main things when she was here in San Diego and I quote: (1) do not work with

PBMCs, they are a mess and (2) put a Bioinformatician in your life… well since I am very creative I will certainly copy the first one but I will modify the second one as: if there flow cytometry involved… get Marta she is the best!!! Nah… I mean she is really the best (and I have learned a lot of flow cytometry with her) but what I meant was that a Bioinformatician is good when the other part is also good and Marta is really good in what she does and super brilliant. Martita, te quiero mucho, y sabes de todo corazón que representas una parte importante en mi vida ya que no creo que hubiese podido llegar (al menos cuerdo) a donde estoy si no hubiese sido por ti. Sabes que siempre he dicho que tu y yo somos un complemento brutal para la investigación y creo que podemos hacer grandes cosas juntos, ya verás que si.

Speaking of research and friends who helped with my sanity I have to mention Topher, George and Sara. (1) Topher, it is funny because I remember one of the first things you taught me was how to be diplomatic when we had to harass people because they were irresponsible or to acknowledge someone no matter what. Every time I have to write an email like that I think even in the words you would use. Also, I think we have done good things together… you got me through my qualifying exam that ended up in my first first-author publication and also the foundations of my genomics knowledge have been because of you, so thanks a lot (and thanks for all the supporting words and gestures in these past 2 months when I have needed them the most)!!!! (2) George, if it wasn't for you probably I would not have been able to graduate at this time. You took really precious time from yourself to help me in all the ways you could and you don't know how much you really helped me, so thanks a lot!!! I hope everything works out awesome in your post-doc!! (3) Sara, what can I say… you have been a great support to me. I don't know how you have done it but interacting with you, knowing you, talking to you and working with you have made me think that my work is very valuable and that I can do anything I want to do! In Davey's absence you have held the fort pretty awesome so I say it is time to TAKE OVER! HASTA LA VICTORIA!

Last, but not least I want to thank Davey. Even though I am always complaining that Davey is not present when I need him the most and he complains that I am too noisy and that he wants to send me far away so he has some tranquility, I think we have an awesome relationship. In reality, even when I complain, sometimes I feel that somehow everything is part of a mystery plan for me to learn something new, which always happens. Davey, has been more than my mentor, he has been like a big brother to me. He has even called me with other names such as “this one”, “that one” or “freezeboy”… but he has always looked out for me as well. I have learned many things from him outside the HIV world that have got me through life here in San Diego. Davey, thanks for giving me the opportunity of working with you. You took me in without knowing absolutely anything and had a lot of patience (especially for the past 2 months) and looking all I have accomplished right now, I have to say that I owe it mostly to you. I am really proud of you and I think you have done a really good job!!! I hope that no matter where I go and what I do, I will always have you there with me! Te quiero un montón!

PS: I will always keep in mind your main life advises: (1) don’t be evil and (2) never do meth! (Maybe it is too late for the first one…)

TABLE OF CONTENTS

Signature Page ………………………………………………………………… iii

Dedication ……………………………………………………………………… iv

Table of Contents …………………………………………………………….... viii

List of Figures …………………………………………………………………. ix

List of Tables ……………………………………………………………......... x

Acknowledgements ………………………………………………………….. xi

Vita …………………………………………………………………………….. xii

Abstract of the Dissertation ……………………………………………....... xiv

Introduction …………………………………………………………………… 1

Study cohort ..………………………………………………………………… 9

Chapter 1 Characterization of the human gut microbiome ………………... 12
1.1 Introduction ………………………………………………………. 13
1.2 Methods …………………………………………………………….. 15
1.3 Results ……………………………………………………………… 19
1.4 Discussion ………………………………………………………….. 21

Chapter 2 The human gut microbiome as a predictor of immune health ... 30
2.1 Introduction ………………………………………………………. 31
2.2 Methods …………………………………………………………….. 33
2.3 Results ……………………………………………………………… 35
2.4 Discussion ………………………………………………………….. 38

Chapter 3 The role of *Lactobacillales* during acute/early HIV infection ….. 53
3.1 Introduction ………………………………………………………. 54
3.2 Methods …………………………………………………………….. 56
3.3 Results ……………………………………………………………… 59
3.4 Discussion ………………………………………………………….. 63

Future Directions ……………………………………………………….…..... 73

Glossary ………………………………………………………………………. 76

References ……………………………………………………….…………… 77

LIST OF FIGURES

Figure 0.1: Gut microbiota targeted therapy 8

Figure 0.2: Microbial translocation model 9

Figure 1.1: Distribution of unclassified sequences in gut bacterial profiles .. 25

Figure 1.2: Comparison of gut bacterial profiles between repeated samples from different PCR amplification 26

Figure 1.3: Gut bacterial profiles during HIV infection 28

Figure 1.4: Distributions of inter- and intra-patient variability 29

Figure 2.1: Unsupervised clustering at baseline 45

Figure 2.2: Bacterial Differences between participants with low vs. high CD4% (Group 1 vs. Group 2) at baseline 46

Figure 2.3: Overview of CD4%, VL and GBP at baseline 47

Figure 2.4: Distribution of CD4% at baseline 48

Figure 2.5: Unsupervised clustering at week 24 of ART 50

Figure 2.6: Distribution of CD4% at week 24 of ART 51

Figure 2.7: Changes in gut flora before and after HIV infection 53

Figure 3.1: Associations of *Lactobacillales* with clinical and immunological variables before ART 68

Figure 3.2: Immune activation at baseline between early vs. acute infected individuals 69

Figure 3.3: Longitudinal associations of *Lactobacillales* with CD4% and soluble CD14 70

Figure 3.4: Associations of *Lactobacillales* with clinical and immunological variables after ART 71

Figure 3.5: Association of immune activation with the HIV latent reservoir .. 72

LIST OF TABLES

Table 0.1: Demographics and clinical variables at baseline ……...…………… 12

Table 1.1: Differences of gut bacterial profiles between re-sequenced samples 27

Table 2.1: Differences in clinical variables between groups demarcated by gut bacterial profiles …..……………………………………………… 44

Table 2.2: Supervised classification statistics ……...…………………………… 49

ACKNOWLEDGEMENTS

All chapters have been submitted for publication. Josué Pérez-Santiago, Sara Gianella, Marta Massanella, Celsa A. Spina, Maile Y. Karris, Susanna R. Var, Derek Patel, Parris S. Jordan, Jason A. Young, Susan J. Little, Douglas D. Richman and Davey M. Smith. Gut *Lactobacillales* are Associated with Higher CD4 and Less Microbial Translocation during HIV Infection. The dissertation author was the primary investigator and author of this material.

VITA

2005	Bachelors of Science, University of Puerto Rico, Rio Piedras Computer Science and Mathematics
2005-2006	Teacher Assistant, University of Puerto Rico, Rio Piedras
2006-2012	Research Assistant, University of California, San Diego
2012	Doctor of Philosophy, University of California, San Diego Bioinformatics and Systems Biology

PUBLICATIONS

1. **Pérez-Santiago, J.**, Diez-Alarcia, R., Callado, L. F., Zhang J. X., Chana G., White, C. H., Glatt, S. J., Tsuang, M. T., Everall, I. P., Meana, J., and Woelk C. H. A combined analysis of microarray gene expression studies of the human prefrontal cortex identifies genes implicated in schizophrenia. Accepted in Journal of Psychiatric Research. (2012).

2. Tilghman, M. W., May, S., **Pérez-Santiago, J.**, Ignacio, C. C., Little, S. J., Richman, D. D., and Smith, D. M. A combined screening platform for HIV treatment failure and resistance. *PLoS One* **7**, e35401 (2012).

3. Woelk, C. H., Singhania, A., **Pérez-Santiago, J.**, Glatt, S. J., and Tsuang, M. T. The utility of gene expression in blood cells for diagnosing neuropsychiatric disorders. *Int Rev Neurobiol* **101**, 41-63 (2011).

4. Woelk, C. H., Beliakova-Bethell, N., Goicoechea, M., Zhao, Y., Du, P., Rought, S. E., Lozach, J., **Pérez-Santiago, J.**, Richman, D. D., Smith, D. M., and Little, S. J. Gene expression before HAART initiation predicts HIV-infected individuals at risk of poor CD4+ T-cell recovery. *AIDS* **24**, 217-222 (2010).

5. Wang, J., **Pérez-Santiago, J.**, Katz, J. E., Mallick, P., and Bandeira, N. Peptide identification from mixture tandem mass spectra. *Mol Cell Proteomics* **9**, 1476-1485 (2010).

6. Tilghman, M. W., Guerena, D. D., Licea, A., **Pérez-Santiago, J.**, Richman, D. D., May, S., and Smith, D. M. Pooled nucleic acid testing to detect antiretroviral treatment failure in Mexico. *J Acquir Immune Defic Syndr* **56**, e70-74 (2010).

7. Manosuthi, W., Butler, D. M., **Pérez-Santiago, J.**, Poon, A. F., Pillai, S. K., Mehta, S. R., Pacold, M. E., Richman, D. D., Pond, S. K., and Smith, D. M. Protease polymorphisms in HIV-1 subtype CRF01_AE represent selection by antiretroviral therapy and host immune pressure. *AIDS* **24**, 411-416 (2010).

8. Smith, D. M., May, S. J., **Pérez-Santiago, J.**, Strain, M. C., Ignacio, C. C., Haubrich, R. H., Richman, D. D., Benson, C. A., and Little, S. J. The use of pooled viral load testing to identify antiretroviral treatment failure. *AIDS* **23**, 2151-2158 (2009).

FIELDS OF STUDY

Major Field: Bioinformatics and Systems Biology

ABSTRACT OF DISSERTATION

The Human Gut Microbiome: Characterization Challenges and HIV Disease Association

by

Josué Pérez Santiago

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2012

Professor Douglas D. Richman, Chair

The human gut microbiome plays an essential role in the maturation and maintenance of the gut associated lymphoid tissue (GALT), yet how the composition of this microbiome assists in this function is not completely understood. Most previous gut flora studies have been limited since they required the isolation and cultivation of microbes, and many of the gut microbes cannot be cultured. The advent of next generation sequencing has improved the identification of the composition of the human gut microbiome, but these techniques require better methods for analyzing these high dimensional data. To overcome these bioinformatics challenges, a pipeline was developed and applied

for the study of the gut microbiome during Human Immunodeficiency Virus (HIV) infection.

The hallmark of HIV infection is the depletion of CD4 T cells, particularly in the gastrointestinal (GI) tract. This depletion compromises gut integrity and results in the translocation of microbial products into the bloodstream. Using next generation sequencing, we found that during untreated HIV infection, gut bacterial profiles segregated 11 participants according to their systemic lymphocyte percentages (Group 1: median CD4% = 9.5% vs. Group 2: median CD4% = 33%). Additionally, regression analyses showed that *Lactobacillales* in the distal gut were associated with higher CD4 counts and CD4%, less immune activation, less viral replication, less gut T cell proliferation and less microbial translocation in untreated and treated HIV infection. Changes of gut bacterial populations may reflect gut immune health and therefore interventions targeted to gut microbiota might help rebuild gut integrity during HIV infection.

Introduction

High throughput: big data

‘Big data’ is a term that has become popular over the past decade with the development of new technologies that allow the generation of enormous amounts of data. According to McKinsey Global Institute, analyzing large data sets will be essential to competition, innovation and productivity in different business and healthcare sectors (http://www.mckinsey.com/). In life sciences, this exponential growth of data has enabled the creation of large data depositories. These data can enable researchers to tackle questions that were previously limited by technology.

One of the first high-throughput technologies that was used in health care research was the microarray. Using microarray technologies researchers had the ability to screen thousands of genes between two or more conditions or genotype regions of a genome to explore whether mutations were associated with disease[1]. This large step enabled a new way of monitoring diseases, examining effects of certain treatments and the discovery of biomarkers that could be used in clinical practice for diagnosis or prognosis of conditions and outcomes[2,3]. However, since nucleic acid expression levels do not necessarily correlate with protein synthesis, or even further post-translational modifications can affect the function of a protein, proteomics became the hot spot. Mass-spectrometry became the high throughput tool for protein quantification[4,5], metabolomics profiling[6], protein-protein interaction[7] and even assessment of spatial molecular

rearrangements of tissue sections[8]. But ultimately, everything is encoded in our genetic material and that is why this is the era of next generation sequencing[9]. With the depth and sensitivity that this technology brings, fields such as metagenomics are being explored. This may represent one of the biggest technological leaps in health care research since diseases and other conditions can be deeply characterized but hypothetically, clinical power of these ‘big data’ technologies have not been realized. Perhaps, we could tailor new, precise and personalized therapeutic interventions.

As generating data becomes cheaper and accumulation of data increases, computational power will become the problem. Consequently, there will be an ever-increasing need of bioinformaticians who can develop new protocols to accurately analyze these ‘big data’.

Host-gut microbiome ecology

One such application of these high throughput technologies is characterizing the human microbiome[10]. Next generation sequencing has the potential to unravel the huge microbial population that inhabits in and on us. The amount of microbes (comprising bacteria, archaeas, viruses and other eukaryotes) that reside in our bodies surpass the number of our ‘own’ cells and are essential to many processes in human life including health and disease. The Human Microbiome Project (HMP) has allowed us to redefine concepts like ‘healthy’ or ‘normal’ according to the diversity in composition of our microbiome. In this thesis, I have focused on the gut microbiome.

The establishment of the gut flora occurs early in life, more specifically during the first three years[11]. The first microbial exposure that infants experience occurs at the time of birth and introduces contact with maternal microbiomes, i.e. vaginal, urogential, skin and gut[12]. For example, the vaginal microbiome is characterized by the presence of distinct *Lactobacillus* species[13]; therefore, infants delivered vaginally tend to have higher proportions of *Lactobacillus* compared to babies delivered by cesarean section[14]. Likewise, other factors including location of delivery (hospital vs. home), time of gestation (prematurity), infant feeding (breast vs. formula), disease and antibiotic use are also involved in what comprises the human gut microbiome[12,15]. After the first three years, the composition of the gut microbiome tends to acquire a ‘stable’ state and, although this ‘stability’ can be a function of geography, genetics, age and diet[11], there is less variation within rather than across individuals[13] and ultimately results in an individual “fingerprint”[16]. One can see this ‘stable’ state more as a functional state since despite the variation in composition and abundance of the gut microbial populations, the metabolic pathways carried by these microorganisms tend to be similar[13] suggesting the idea of functional redundancy.

The gut microbiome is essential to maintain good health since apart from processes such as metabolism of nutrients and xenobiotics and tissue differentiation, they are very important for the development, maturation and assistance of our immune health. For example, gut microbiota can directly interact with pathogens by direct inhibition (for example secreting metabolites and bacteroicins) or nutrient depletion to prevent colonization[17,18]. Likewise,

different gut microbial species can regulate different types of immune cells (for example *Lactobacillus* and *Bifidobacterium* regulate Tregs) and also modulate various immune responses (for example the production of anti-microbial peptides)[19]. Further, alteration to the 'stable' state of the gut flora is associated with diseases such as cancer, allergies, diabetes, Crohn's disease among others[19-23]; therefore, taken together with recent discoveries, the idea of manipulating the gut flora to treat disease states has become more tempting[24]. One can visualize the dynamics of the gut flora as states (representing the community composition of the gut flora) in a landscape with basins (representing the likelihood of changing states), where changing from one 'stable' state to another 'stable' state require major disturbances (**Figure 0.1**). Consequently, the goal and major challenge of microbiota-aimed therapeutics is to decipher how to climb these basins to ultimately move from a "disease" state to a "healthy" state perhaps by the intake of prebiotics probiotics and/or symbiotics, antibiotic regimen, change in diet and fecal transplants among others.

HIV pathogenesis

Human immunodeficiency virus (HIV) is one of the biggest infectious burdens worldwide. At the end of 2010, it was estimated that 34 million people were living with HIV, with 2.7 million new infections per year (http://www.who.int/). Being a retrovirus, HIV integrates into the host genome thereby allowing it to become dormant and difficult for the immune system to target and eradicate. Acute HIV infection starts with high level of circulating virus

in the host resulting in high levels of immune activation. During the chronic phase, immune activation continues and eventually exhausts the T-cell immune system and leads to Acquire Immunodeficiency Syndrome (AIDS), with the appearance of opportunistic infections and ultimately death.

Together with the loss of peripheral CD4 T-cell counts in HIV-infected individuals, there is a rapid and enormous depletion of CD4 T cells particularly in the gut associated lymphoid tissue (GALT)[25-27], which contain the majority of T cells in the body[28] with the majority of them expressing CCR5. The GALT maintains the integrity in the gastrointestinal (GI) tract acting as a barrier for microbes. Therefore, depletion of the CD4 T-lymphocytes in the GALT increases gut permeability with microbial products, such as LPS, translocating into the bloodstream. This translocation of microbial products elicits a systemic inflammatory response, resulting in the presence of higher plasma levels of inflammatory cytokines and thus activation of monocytes, macrophages and lymphocytes. This chronic systemic immune activation provides the virus a larger source of cells with an activated phenotype that can be infected accelerating the progression to AIDS (**Figure 0.2**).

Simian immunodeficiency virus (SIV) infection in its natural hosts, e.g. sooty mangabeys and African green monkeys, is asymptomatic, does not induce a depletion of CD4 T cells in the majority of animals and does not progress to AIDS. Whereas depletion of CD4 T cells in the gut occurs in SIV and HIV infection[25-27,29,30], there is no evidence of microbial translocation or chronic immune activation in natural host of SIV[31]. It has been suggested that natural

hosts of SIV infection preserve GALT functionality, possibly by maintaining the balance of specific T-cell subsets (for example Th17) despite the massive CD4 T-cell depletion[31]. GALT depletion results in microbial translocation that is associated with chronic immune activation and hence with HIV disease progression[32-34]

Antiretroviral therapy (ART) often controls viral replication and results in the restoration of CD4 T cells, especially if it is initiated early (within the first months of infection)[26]. However, the effects on the GI tract are poorly understood; only some individuals may partially restore the GALT and reduce microbial translocation[26,35,36]. Overall, the physiologic GI microbiota help maintain gut integrity decreasing the likelihood of microbial translocation, but during HIV infection the gut microbiome is disturbed[37]. Furthermore, during HIV infection, the consumption of probiotics is not only associated with an increase of CD4 T cells[38,39] but also with less immune activation[40]; therefore understanding the underlying involvement of the gut flora during HIV infection is crucial to prevent or treat HIV.

To my knowledge, there is no study that has associated gut microbial populations with HIV pathogenesis. Therefore in this thesis, I followed 13 participants before and after the initiation of ART with the objective of characterizing the changes in the GI bacterial flora of HIV infected as they progress through acute, early and chronic phases of infection and if/how ART modulates these changes. In the process, I approach the most common

experimental and bioinformatics challenges in bacterial metagenomics studies and a general pipeline that alleviates these challenges.

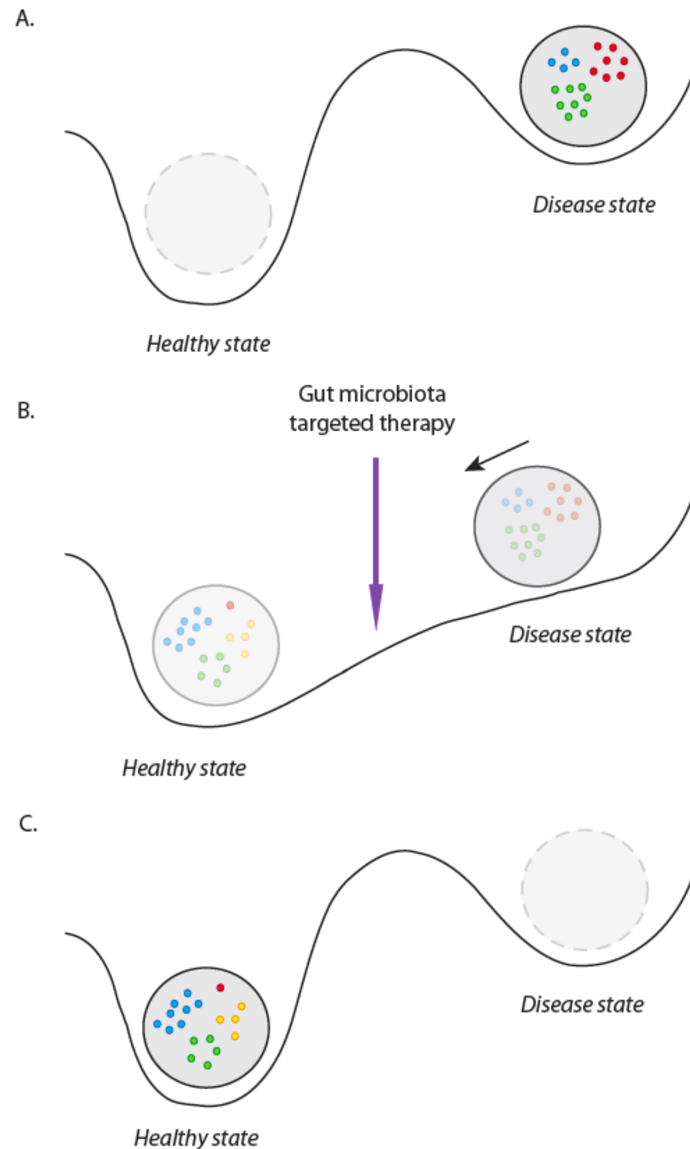

Figure 0.1: Gut microbiota-targeted therapy

(A) The objective of gut microbiota-targeted therapy is to change the composition of the gut flora to a healthier state. Circles represent stable states of the gut flora and smaller colored circles are the different bacterial populations that compose the stable state of the gut flora. Circles are in a landscape with basins where the depth of the basins represents the likelihood that the composition of a state remains unchanged. The slopes represent the amount of perturbation needed to change from one stable state to another in the landscape, by lowering the slopes (B). To move from a disease state, a major disturbance is needed in order to the circle to move permanently to reach a healthy state (C).

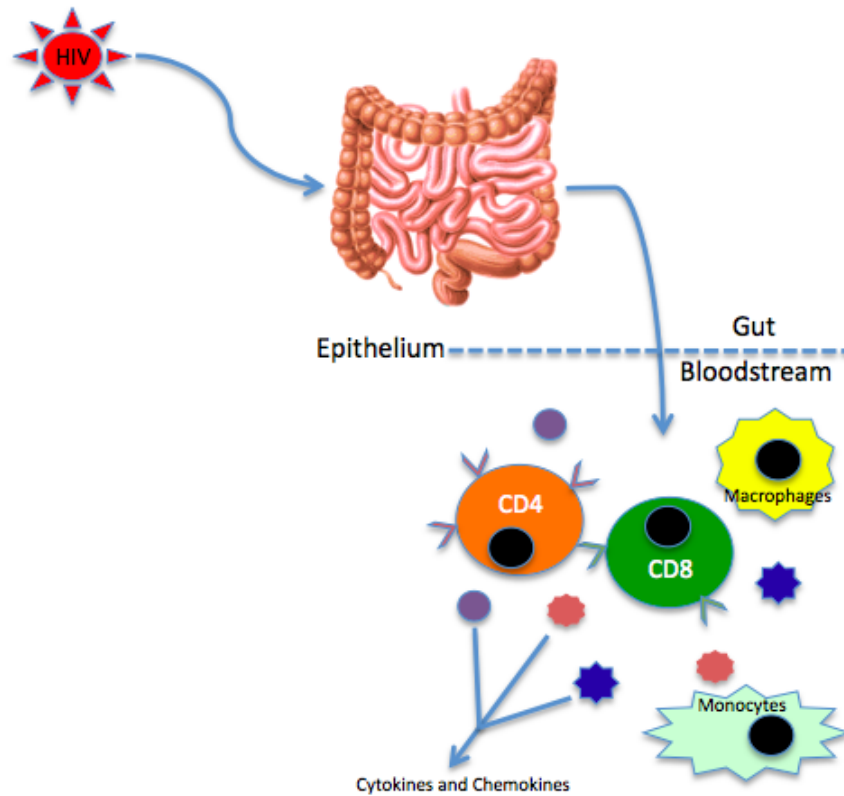

Figure 0.2: Microbial translocation model

After infection of HIV a huge depletion of CD4 T cells occurs in the body. Particularly, this depletion happens faster in the gut associated lymphoid tissue (GALT), which protects the body from external pathogens and maintains integrity and equilibrium of commensal microbes living the gastrointestinal (GI) tract. This allows a breakdown of epithelial cells resulting in the migration of these microbes and microbial products into the bloodstream activating the immune system and hence causing systemic inflammation.

Study cohort

Eligible participants were men who had sex with men (MSM) enrolled in the San Diego Primary Infection Cohort (N=13) and a randomized double blind control ART trial of maraviroc vs. placebo. The study cohort was comprised of 10 Whites and 3 Asians with an average age of 33 years and estimated duration of infection (EDI) of 6.5 weeks. The Institutional Review Board of our center approved this study and all participants provided written informed consent. All patients initiated ART within 2 weeks of study enrollment with a combination of tenofovir, emtricitabine and ritonavir-boosted atazanavir, with or without maraviroc. The double-blind clinical trial is ongoing with all patients remaining blinded to maraviroc use. Anal swabs, blood, semen, peripheral lymphocyte profiles, and HIV levels (Amplicor, Roche) were collected at baseline (within a week before the initiation of ART) and approximately every 4 weeks thereafter for 48 weeks. Epidemiological, behavioral risk and HIV-related data were also collected from the participants. We determined EDI using results of serologic and virologic tests as described previously[41]. A summary of clinical variables measured at baseline is provided in **Table 0.1.**

Gut biopsies were available for a subgroup of nine patients at weeks 0 and 48 and interval biopsies at weeks 12 or 24 for a proportion of the subgroup. These participants underwent colonoscopies during which nine mucosal biopsies were obtained at rectosigmoid junction and terminal ileum. Tissue samples were incubated with collagenase and DNAse prior to passage through a cellular

strainer (PGC Scientifics) and re-suspension in freeze medium (90% heat inactivated fetal bovine serum and 10% DMSO) before storage at -140°C.

Table 0.1: Demographics and clinical variables at baseline

PID	Race[1]	Ethnicity[2]	Age	EDI (weeks)	Viral Load (log_{10} copies/mL)	CD4 Absolute (cells/μL)	CD4 %	CD8 Absolute (cells/μL)	CD8 %	CD4/CD8 Ratio
A	A	NH	23	2	5.78	276	10	1890	70	0.15
B	A	NH	28	2	3.10	493	24	856	45	0.58
C	W	NH	35	3	6.78	525	9	5216	85	0.10
D	W	H	22	3	6.99	91	9	728	70	0.13
E	A/W	NH	39	3	4.15	692	29	1387	55	0.50
F	W	NH	52	3	5.36	374	28	732	53	0.51
G	W	NH	21	3	5.93	916	37	1082	44	0.85
H	W	NH	28	4	7.00	520	11	3334	71	0.16
I	W	H	28	10	3.89	501	33	775	51	0.65
J	W	NH	26	11	5.25	614	33	776	42	0.79
K	W	H	33	12	4.40	971	34	1086	38	0.89
L	W	NH	40	14	4.15	383	30	558	42	0.69
M	W	NH	55	14	3.77	913	40	714	31	1.28
Average			**33.1**	**6.46**	**5.12**	**559.15**	**25.15**	**1471.85**	**53.62**	**0.56**

[1] W = White and A = Asian. [2] H=Hispanic and NH=non-Hispanic. Participants are ordered according to estimated duration of infection (EDI).

Chapter 1

Classification of the human gut microbiome during HIV

Abstract

Next generation sequencing provides improved methods for the identification and classification of microorganisms. However, generation and analysis of these high dimensional data require meticulous study designs and substantial computational power. In the current study, we provide solutions to common experimental issues and a pipeline for the analysis of high throughput data for metagenomics studies. When applied during Human Immunodeficiency Virus (HIV) infection, we were able to classify the human gut microbiome and found that variability of the gut flora is higher across individuals.

Introduction

It is estimated that the biosphere consists approximately of 10^{30} microbes, which play essential roles in all life, varying from environmental balance to human homeostasis and health[42]. However, the composition and functional underpinnings of these microbiomes are not completely characterized. Traditional methods of studying microorganisms required isolation and cultivation or cloning techniques that frequently represented a challenge since most of these microbes cannot be cultured. The advent of next generation sequencing permits the investigations of these microbial populations as thousands of sequences can be obtained from a single biological sample in less time than traditional procedures[43]. For this, Chen[44] redefined metagenomics as the application of modern genomics technique to study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species.

Bacterial metagenomics studies often utilize the 16s ribosomal gene (16s rDNA), consisting of 9 hypervariable regions[45] (denoted as V1-V9), to produce a diversity signature of the biological sample of interest. While this gene is variable across species, it is also well conserved within species. Particularly, regions V2, V3 and V6 have been shown to be the most adequate regions to screen since can distinguish most bacteria[46]. However, a combination of regions would be needed for accurate estimation of species richness of the population[47]. Subsequently, PCR amplification followed by next generation sequencing of

these regions of the 16s rDNA are the next steps for characterizing microbial populations.

Classification of microbial populations is one of the biggest challenges in metagenomics studies. The typical workflow of classification, denoted as taxonomic dependent analysis, results in the assignment of a class by comparing sequences against reference databases. Assignment is often limited by incompleteness of sequence annotation. On the other hand, taxonomic independent analysis assigns operational taxonomic units (OTU) according to sequence variation, which is often used for sequence annotation. While the former methods are powerful estimating richness of microbial populations, some kind of reference is needed to complete annotation. In the present study, we combined both types of analyses to classify and compare the human gut microbiome of HIV infected individuals; nevertheless this pipeline could be applied to any other microbial population.

Methods

Stool DNA isolation

Stool DNA was extracted from anal swabs using the QIAamp Stool DNA kit (Qiagen, CA). Anal swabs were suspended in the ASL buffer for 5 minutes. All remaining steps were followed per manufacturer protocol except that the elution was performed in 200 L incubated for 5 minutes before the final spin to increase DNA yield.

Amplification of bacterial DNA

Amplification of the V6 hypervariable region of the 16s rDNA gene was performed in a 50 L reaction. To reduce contamination coming from the carry over DNA of the polymerase, we used the highly purified Amplitaq Gold Low DNA polymerase (Applied Biosystem, CA). Master mix consisted of 26 L of nuclease free water, 5 L of 10X PCR Gold buffer, 5 L of MgCl2 (25mM), 1   L of deoxynucloside triphosphate (10mM), 1.25 L of primers (20 M, forward primer 967F: 5'-CAACGCGAAGAACCTTACC-3' and reverse primer 1064R 5'-CGACAGCCATGCANCACCT-3'), 0.5 L of Amplitaq Gold LD polymerase and 10uL of sample DNA. The reaction used the following cycling conditions: (1) initial activation at 93°C for 15 minutes, (2) 30 cycles of denaturation at 95°C for 30 seconds, annealing at 57°C for 30 seconds and extension at 72°C for 1 minute, (3) followed by a final extension at 72°C for 10 minutes. Samples were

run in duplicates and 1% agarose gel electrophoresis was used to confirm the ~97bp size of product. Duplicate samples were combined and purified immediately after reaction (Qiagen PCR Purification Kit). Quantification and purity of DNA was assessed using the Agilent 2100 BioAnalyzer.

Pyrosequencing and quality control

Adaptor sequences were linked to purified product (Roche Rapid Library Prep Kit) using 240ng of DNA as starting material. Quantification of library prep was done using the TBS-380 Fluorometer (Turner Biosystems) and a 1:10 dilution of the original elution. Emulsion PCR was carried out using a dilution of 2 copies/bead and pyrosequencing was performed (Roche 454 FLX Titanium). Bacterial sequences with at least 90 continuous base pairs, which contained a quality score of at least 20[48-50], were considered for metagenomics analyses.

Classification of bacteria

Primarily, we classified sequences using the Ribosomal Database Project (RDP)[51] to the order level. Briefly, for each known genera, the frequencies of all possible eight base sub-sequences of the 16s rDNA are calculated and used as training set. A query sequence is assigned to a hierarchical level when it maximizes the probability of observing all subsequences of a particular genus. Classification was performed at the order level and sequences that did not achieve confidence of classification were assigned to OTU based on genetic distance using the tool ESPRIT[52]. ESPRIT uses the Needleman-Wunsch to

calculate the pairwise distance of all sequences followed by hierarchical clustering optimized for handling data in high orders of magnitude. With a genetic variation of 10%, a consensus sequence was built for each OTU and then classified using small subunit rRNA taxonomy and alignment pipeline (STAP)[53]. In short, STAP uses BLAST to select sequences related to the query sequence and build a maximum likelihood tree to make an initial assignment. Using this assignment, it creates a mini database of sequences of the same taxonomic group and constructs another phylogenetic tree to give a final classification. In the current study, gut bacterial sequences were classified as orders of bacteria common across all samples. Sequences that were unique to an individual or did not achieve classification at the order level were categorized as "Other".

Measurement of similarity of bacterial profiles

Bacterial profiles can be modeled as a vector in R^n such that $\vec{x} = (x_1, x_2, ..., x_n)$, where *n* is the number of different classifications (i.e. different phyla or orders of bacteria) and x_i corresponds to the relative abundance of these classifications, hence $0 \le x_i \le 100$ for all x_i and $\sum_{i=1}^{n} x_i = 100$. In Euclidean geometry the dot product between two unit vectors, and , is defined as:

$$x \bullet y = \cos(\theta)$$

where θ is the angle between the two vectors. This angle θ, defined as, $\theta = \cos^{-1}(x \bullet y)$, is used as a measurement of similarity between bacterial profiles as it captures the difference in direction of both vectors. Two vectors are most

similar as the angle approaches zero and they are most different as the angle tends to 180°. The percentage of difference between two vectors is calculated as the ratio of the angle between them divided by 360 (degrees in a circle).

Variability of bacterial profiles and statistical analyses

The sample variance of bacterial profiles was calculated generalizing the sample variance formula for vector calculations. Intra-patient variability was measured calculating the sample variance of the gut bacterial profiles at all time points for that particular patient. Inter-patient variability was computed taking into account all the samples at each time point available. All statistical analyses were performed using R statistical language.

Results

Using RDP and after filtering, we classified a total of 2,266,083 sequences (26,977 on average per sample) representative of distal gut bacterial flora at the order level using V6 region of the bacterial 16s rDNA. Overall, the proportion of unclassified sequences fluctuated across all samples (range: 3.52%-51.1%, **Figure 1.1**). Since classification could be affected by PCR inefficiency or bias, ten samples were re-amplified, re-sequenced and re-classified. There was no statistical difference in the proportion of unclassified sequences between the two repeated samples (two-tailed Wilcoxon: P=0.49). Additionally, comparison between the paired bacterial profiles revealed that differences between profiles ranged from 1.65%-12.41% (**Figure 1.2**). We observed that variation was influenced mainly by the order of *Bacteroidales*. On the other hand, since sequencing errors can also influence classification, we re-sequenced eight samples using the same PCR amplification and all repeated samples showed less than 2% of difference (**Table 1.1**). From the repeated samples, we selected the ones with the highest number of sequences for further analysis.

To alleviate the classification problem, we used the combination of ESPRIT and STAP to classify orders of bacteria based on genetic similarity. Participants shared 14 distinct orders of bacteria and individual orders of bacteria were classified as “Other”. However, the order of *Burkholderiales* showed an increase in proportion associated to the time of the DNA extraction, which may be associated with evidence of contamination. Therefore, to be conservative, we

excluded all sequences from this order and re-calculated relative abundance. A summary of all bacterial profiles is shown in **Figure 1.3**. Overall, the median intra-patient variability in GBP was significantly lower than inter-patient variability (two-tailed Mann-Whitney, *P*=0.006, **Figure 1.4**), as reported previously[13].

Discussion

While next generation sequencing has overcome many challenges in traditional of metagenomics studies, it also has brought new ones. First, inappropriate selection of PCR reagents and thermocycling conditions could lead to contaminant bacterial amplification as most *Taq* polymerases carry bacterial DNA. To reduce this bias, we used a highly purified polymerase (Amplitaq Gold LD, Applied Biosystems, Inc.) and controlled the number of cycles in the reaction (≤30 cycles). Second, PCR bias and errors may have more impact in classification than sequencing. In our case, it may seem that the order of *Bacteroidales* had preferential PCR amplification but it should be noted that variation in re-amplified samples could also be a consequence of downstream processing by 454 methodology. For example, for both samples from Patient H (12.41% and 4.59% of variation), adaptor sequences library preparation failed and had to be repeated, hence PCR product had to be thawed twice possibly resulting in DNA degradation. Additionally, emulsion PCR from Patient F at week 24 (6.59% of variation) did not achieve a good bead recovery (<5%) resulting in a non-optimal pyrosequencing (~10,000 reads). Moreover, samples were re-amplified at time points very far apart, increasing the likelihood of DNA degradation and possibly resulting in the decrease or loss of low abundant bacteria. Third, the proportion of unclassified sequences that can be obtained from a biological sample can be variable[54-56] and current methods ultimately require sequence information from “known” organisms. In other words, we have

to know what we are looking for to classify it. For example, during HIV infection, we found that this unclassified proportion could be more than 50% in a sample. This represented a limitation for correlation of the human gut microbiome and HIV disease progression. With the combination of taxonomic dependent and independent analysis, we were able to reduce this limitation. Although some of these sequences achieved classification at higher levels (kingdom, phylum and class), some remained unclassified at the order level and likely represent novel bacteria. Nevertheless, with the characterization of the human gut microbiome in HIV infection, we also found[13] that variability of the microbial populations is less in individuals over time than across individuals. Fourth, high throughput classification of bacteria presents a computational challenge. For example, STAP can process one sequence in 1 minute and 35.4 seconds on average[53], so processing a 365,043 sequences (amount of unclassified sequences on our study) would take ~403 days. Using ESPRIT, we were able to reduce this order of magnitude to 4,367 representative sequences, thereby using only 2.5 days of computational time.

The massive generation of data permits the characterization of microbial diversity in different habitats but also allows for the discovery of new organisms[43]. While the former could be feasible for association of microbial communities with many kinds of biological processes, such as human disease or carbon fixation[42], etc., the latter may contain information needed to improve these association studies. Selection of the type of workflow will be highly dependent on the question we desire to answer. In HIV research, little is known about the

interaction between the gut flora, HIV and the host, hence our primary research aim was to determine relationships between HIV pathogenesis with previous knowledge about the human gut reducing the amount of unclassified sequences. Therefore, we applied a metagenomics pipeline that although we used for classification of the human gut microbiome in HIV infection, it could be utilized in other metagenomics studies.

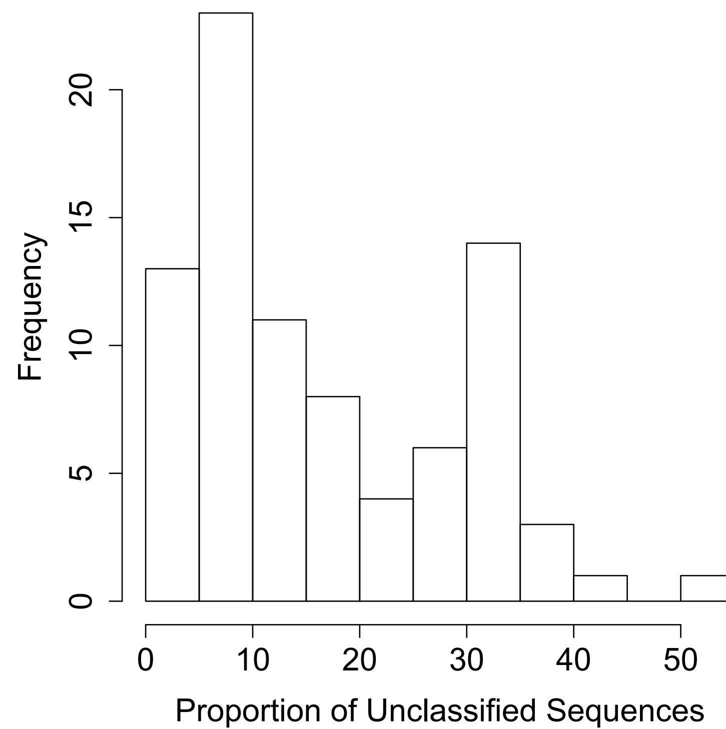

Figure 1.1: Distribution of unclassified sequences in gut bacterial profiles

Gut bacterial sequences were classified at the order level using RDP. The distribution of the relative proportion of unclassified sequences showed a wide range with a maximum of 51.1%.

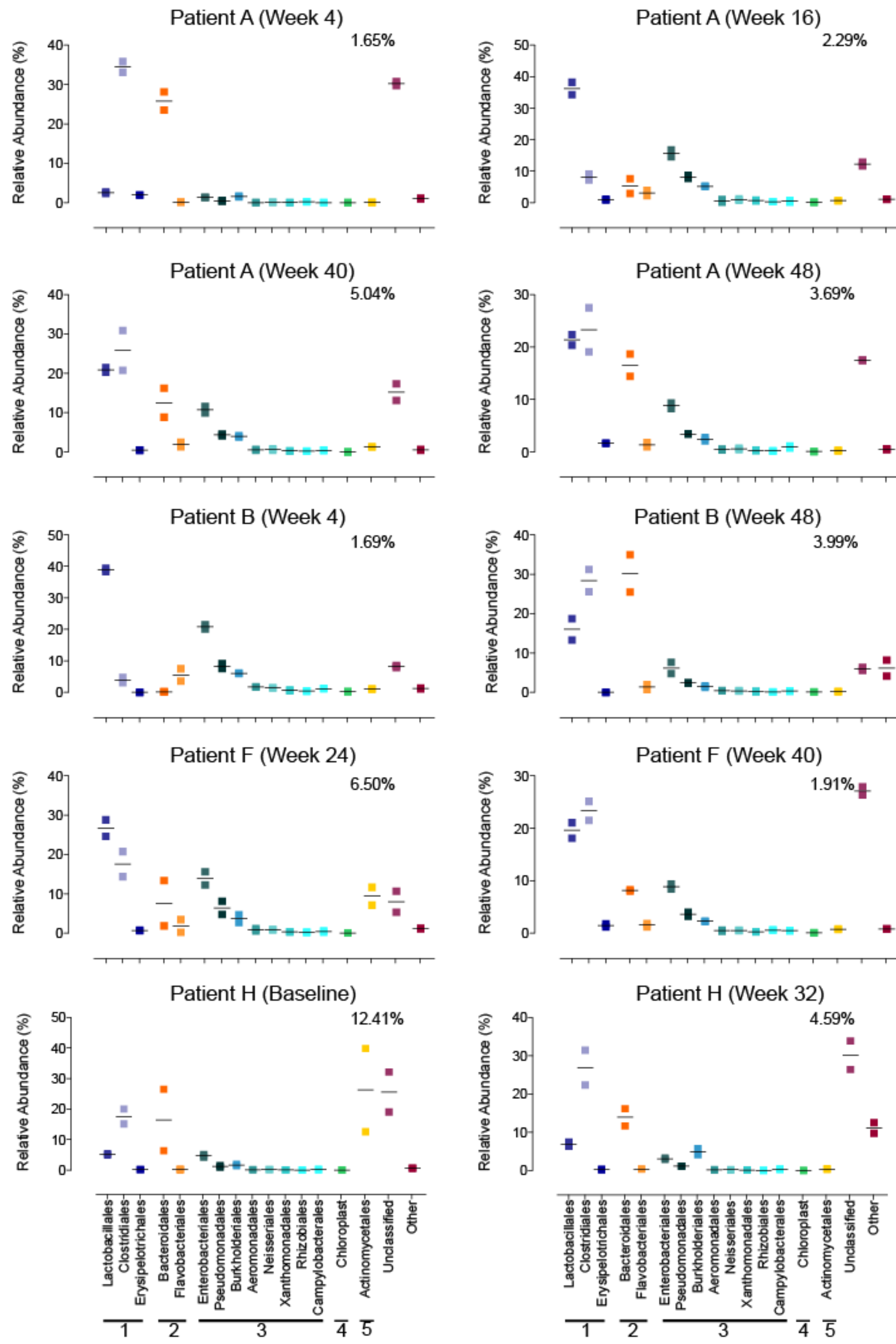

Figure 1.2: Comparison of gut bacterial profiles between repeated samples from different PCR amplification

Using the angle measurement, we calculated the difference between each pair of repeated samples. Bacteria are grouped by phylum and percentage of difference is displayed at the top right of each graph.

Table 1.1: Differences of gut bacterial profiles between re-sequenced samples

Using the angle measurement, we calculated the difference between each pair of repeated samples. BL represents baseline sample. All samples showed less than 2% of difference in classification.

Patient	Time point (week)	Difference (%)
A	BL	0.56
A	8	0.64
A	32	0.43
I	BL	0.63
I	24	0.32
J	BL	0.86
J	24	1.90
H	24	1.29

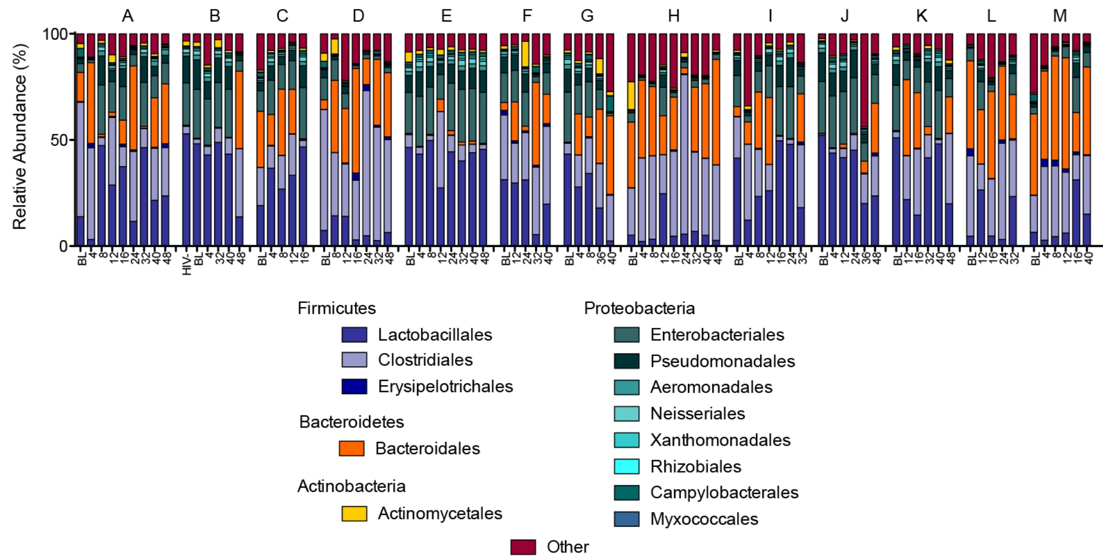

Figure 1.3: Gut bacterial profiles during HIV infection

Using a combination of taxonomic depend and independent classification we classified the gut microbiota of 13 HIV-infected individuals. Baseline (BL) sample was collected within a week of initiation of antiretroviral therapy (ART). Numbers represent the weeks on ART. Orders of bacteria are grouped by phyla and orders in the same phyla share similar colors.

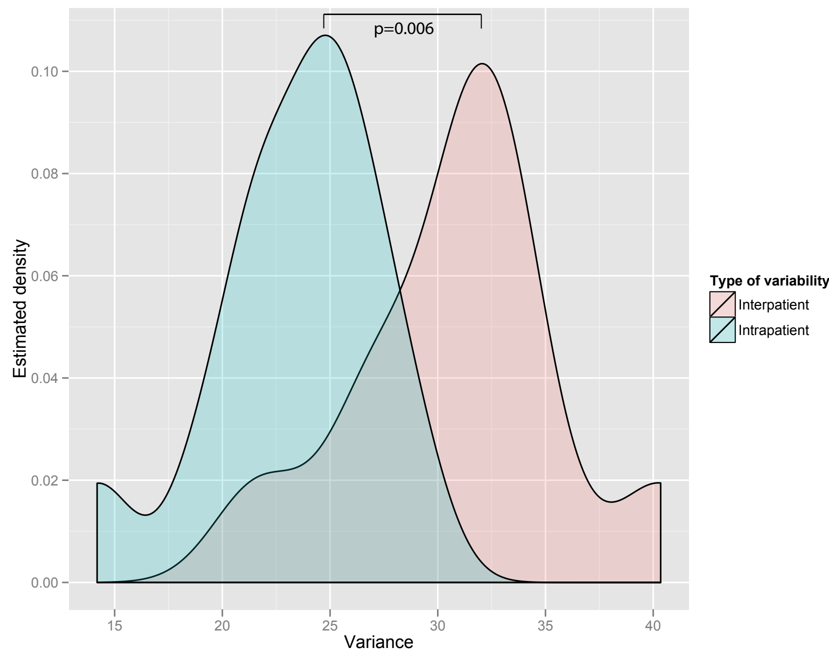

Figure 1.4: Distributions of inter- and intra-patient variability

For each participant, intra-patient variability was calculated utilizing all samples from the same individuals. For each time point sampled, inter-patient variability was measured using all samples at a respective time point. Intra-patient variability (median=24.2) was significantly less than inter-patient variability (median=31.5) when subjected to a Mann-Whitney test. The estimated density functions for inter- and intra- patient variability are shown in red and green, respectively.

ACKNOWLEDGEMENTS

Chapter 1, in part, has been included in the submission for publication. Josué Pérez-Santiago, Sara Gianella, Marta Massanella, Celsa A. Spina, Maile Y. Karris, Susanna R. Var, Derek Patel, Parris S. Jordan, Jason A. Young, Susan J. Little, Douglas D. Richman and Davey M. Smith. Gut *Lactobacillales* are Associated with Higher CD4 and Less Microbial Translocation during HIV Infection. The dissertation author was the primary investigator and author of this material.

Chapter 2

The gut flora as a predictor of immune status during HIV infection

Abstract

Changes in the compositional states of the gut microbiome are associated with health and disease and represent a new source of biomarkers that can be assayed by next generation sequencing. Previously, there has have been no studies have used the gut flora to construct classifiers for clinical purposes. This study identified that different compositional states of the gut flora are associated with different lymphocyte percentages (Group 1: median CD4% = 9.5% vs. Group 2: median CD4% = 33%) during untreated early/acute HIV in 11 individuals. During antiretroviral therapy, we used support vector machines to construct a diagnostic classifier able to determine participants' current immune status in blood (Group 1 vs. Group 2). Although classification achieved a 88% accuracy, these changes in bacterial populations possibly reflect gut immune health. This represents a novel screening intervention for gut immune recovery during HIV infection.

Introduction

High throughput technologies have facilitated the development of clinically relevant biomarkers[57,58]. For example, gene expression in peripheral mononuclear blood cells (PBMCs) have been utilized for the construction of gene expression classifiers able to diagnose disease states, predict disease outcome and determine a patient's response to drug treatment[3]. The current era of next generation sequencing has enabled pioneering studies to better characterize the composition, functionality and factors defining the healthy human-gut ecosystem[10,11,13], which represents a new source for the development of biomarkers.

Disturbances of the gut microbiota often result in a specific shift of microbial populations[20,23]. During early Human Immunodeficiency Virus (HIV) infection there is a rapid and preferential depletion of CD4 T cells in the gastrointestinal (GI) tract[25-27] and subsequent shifts in the gut flora might occur[37]. While antiretroviral therapy (ART) often results in a decline in viral replication and recovery of systemic CD4 T cells, the effects on the GI tract remain unclear. The objective of this chapter is to investigate how differences in lymphocyte profiles might impact the gut flora populations in the absence of ART and the possible use of these shifts in bacterial populations to define immune health after the initiation of therapy. Gut bacterial profiles (GBP) of 13 HIV+ individuals were evaluated within a week of starting ART and approximately every 4 weeks after the initiation of ART for 48 weeks. Immune health (lymphocyte proportions) in

relation to gut flora was determined at baseline and supervised machine learning methods were used to predict immune status after the initiation of ART.

Methods

Gut bacterial profiles and clinical variables

Classification of bacteria, described in the previous chapter, was used for the current analysis. Briefly, DNA extracted from anal swabs was amplified and pyrosequenced for the V6 region of the bacterial 16s rDNA and classified bacteria at the order level. Samples were collected shortly before the initiation of therapy (baseline, n=13) and approximately every 4 weeks for 48 weeks (n=70). Additionally, peripheral lymphocyte profiles and HIV levels (Amplicor, Roche) were measured at the same time points. For one participant an anal swab was collected prior to HIV infection. Additionally, since antibiotics impact gut flora, two participants (L and M) were excluded from baseline analysis because antibiotics were used shortly before the collection of their baseline samples.

Microbial translocation markers

Two markers of bacterial translocation, lipopolysaccharide (LPS) and Soluble CD14 (sCD14), were measured in blood plasma following manufacturer's protocols (Limulus amebocyte Lysate QCL-1000 for LPS and Quantikine ELISA Human sCD14 Immunoassay for sCD14).

Unsupervised clustering and statistical tests

Immunologic states were identified using unsupervised clustering on the basis of participants GBP and a two-sided Mann-Whitney statistical test was

used to assess differences between states. All statistical analyses were performed using R statistical software.

Classification algorithm

To investigate whether GBP could be used to assign HIV-infected individuals to a group (Group 1: median CD4% = 10.6% vs. Group 2: median CD4% = 46.5%)based on their lymphocyte profiles, we used support vector machines (SVM) since they have been applied in a variety of biomedical applications with different types of high throughput data[59]. Given the training data, the SVM constructs a hyperplane in a multidimensional space that maximizes the distance between the nearest members of the two classes, minimizing the error of classification. To avoid overfitting and obtain the most informative features, methods of regularization were applied. Therefore, we constructed a SVM based on the orders of *Lactobacillales*, *Bacteroidales* and *Rhizobiales* (SVM_{LBR}) using the package LiblineaR in R. Validation of SVM_{LBR} was assessed using leave-one-out-cross-validation (LOOCV).

Results

Class discovery using gut bacterial profiles

To avoid assumptions of associations of GBP with clinical variables, an unsupervised clustering analysis of baseline orders of bacteria was performed and revealed two distinct GBP profiles (**Figure 2.1**). The first GBP cluster (Group 1) compared to the second GBP cluster (Group 2) showed significantly lower proportions of *Lactobacillales* (*P*=0.006, two-tailed Mann-Whitney), *Enterobacteriales* (*P*=0.006), *Pseudomonadales* (*P*=0.006), *Xanthomonadales* (*P*=0.006), *Aeromonadales* (*P*=0.006), *Rhizobiales* (*P*=0.006) and *Neisseriales* (*P*=0.006), and higher proportions of *Bacteroidales* (*P*=0.01) and *Clostridiales* (*P*=0.04) in the distal gut (**Figure 2.1** and **2.2**). Interestingly, participants in these two GBP groups differed also significantly by the percentage of CD4 (CD4%) and CD8 (CD8%), and viral load (VL) at baseline, but there were no differences in markers of microbial translocation or estimated duration of infection (EDI, **Table 2.1**). While Group 1 had low CD4% (median = 9.5%), high CD8% (median = 70.5%) and high VL (median = 6.90 HIV RNA $\log_{10}$ copies/mL), Group 2 had high CD4% (median = 33%), low CD8 (median = 45) and low VL (median 4.39 HIV RNA $\log_{10}$ copies/mL, *P*=0.01, for all parameters, **Table 2.1** and **Figure 2.3**). The more profound differences between groups were observed in the proportion of *Lactobacillales*. Group 1 with low CD4%, high CD8% and high VL had low *Lactobacillales*, while Group 2 with high CD4%, low CD8% and low VL had high *Lactobacillales* (median 10.6% vs. 46.5%, *P*=0.006, **Figure 2.2** and **2.3**). Since

the proportion of gut *Lactobacillales* has been associated with how a person is born[14], we also examined reported differences between vaginal versus cesarean delivery of participants and found no differences between Groups 1 and 2 (Fisher exact test, *P*=0.48).

Gut bacterial profiles as predictor of immune status

The ability of using GBP to predict a participant's immune status (Group 1: median CD4% = 9.5% vs. Group 2: median CD4% = 33%) after the initiation of therapy was assessed using SVM_{LBR} that was developed from analysis of baseline GBP. Briefly, the SVM_{LBR} was built using methods or regularization with the orders of bacteria that were significantly different between groups. Using LOOCV, SVM_{LBR} was validated and exhibited a 100% accuracy discriminating Group 1 vs. Group 2 when trained with baseline samples. Since both groups showed distinct CD4% distributions at baseline (**Figure 2.4**), we tested intermediate values of CD4% between the two distributions to find the value of CD4% that maximized classification. The maximum accuracy value obtained was 88% at values of CD4% of 14%-16% with a sensitivity of 0.75, specificity of 1, a positive predictive value of 1 and negative predictive value of 0.06 (**Table 2.2**). A low sensitivity and negative predictive value are associated with a high rate of false negatives likely meaning that individuals had a compositional state similar to those individuals with low CD4% (Group 1) but they were still classified as high CD4% (Group 2). Interestingly, using unsupervised clustering after 24 weeks of ART, participants in Group 1 could still be segregated by their GBP from

participants in Group 2 (**Figure 2.5**). Also at week 24, Group 1 individuals still had overall lower CD4%, but the clear demarcation in CD4% observed at baseline was no longer observed, probably because all participant's increased their CD4% during ART (**Figure 2.6**).

Since CD8% was significantly different in between groups, we defined immune status based on CD8%. However, GBP failed to predict whether a sample belongs to Group 1 vs. Group 2 (data not shown). On the other hand, VL was not used for predictions since after initiation of ART VL declines to undetectable levels.

Gut bacterial profiles before and after HIV infection

For one of our participants (Patient B), we had one available anal swab prior to HIV infection. There was no compositional change in GBP before and after HIV infection as both profiles exhibited a 1.5% of difference (angle measurement defined in chapter 1, **Figure 2.7**). Further, SVM_{LBR} assigned him to Group 2 before his HIV infection, which remained after infection. Even though there was not a lymphocyte profile available for him prior to infection, at baseline he had 28% of CD4 (374 cells/μL), which also classified him as Group 2. Taken together, this suggests that Patient B had GBP consistent with higher CD4% even before his HIV infection.

Discussion

Other groups have encountered that various disease states can be classified by specific shifts of the bacterial populations residing in the gut using unsupervised methods of machine learning[20,23]. To our knowledge, none of these studies have used supervised methods, such as SVMs, to translate this information into clinical practice. The goal of unsupervised learning is to determine whether groups can be found based on similarity of pieces of data and not any class label; i.e. it clusters similar data together and it assigns a label (class discovery). On the other hand, supervised methods of machine learning use the predetermination of classes to find patterns in the data in order to build a mathematical function for predictive purposes, i.e. map predictors to classes. Using unsupervised learning, two different groups of patients were identified by their GBP that can be associated with different levels of CD4% (and also CD8% and VL) during non-treated HIV infection. With two classes defined by CD4%, we applied supervised learning on GBP to construct a diagnostic classifier able to predict an individual's immune state (CD4%) after the initiation of ART.

The hallmark of HIV infection is the vast depletion of systemic and preferential gut CD4 T-cells and consequently the gut integrity is compromised. Initiation of ART blocks viral replication and promotes recovery of CD4 T-cells, although this recovery varies across individuals and between anatomical compartments[26,35,36]. In other words, in acute and untreated HIV infection both compartments (blood and gut) are similarly damaged but the impact of ART on

immune reconstitution may differ between compartments and thereby shifts in GBP might be mainly reflecting gut immune status rather than lymphocyte status in the blood. Several of our study's results support this theory. First, our classifier achieved an 88% accuracy predicting whether participants had high vs. low CD4%. However, errors in classification were primarily because of false negatives. In other words, after ART participants were recovering their CD4 T-cells as expected, but shifts of the gut flora remained representative of participants with low CD4%. Second, although there was no statistical difference in microbial translocation markers (LPS and sCD14) between groups at baseline, sCD14 was higher in Group 1, almost achieving significant levels (P=0.07). Third, using unsupervised clustering after 24 weeks of ART, there were still two main clusters and although participants with lower CD4% remained in the same cluster, all participants had increased their CD4%. Although not conclusive, it is tempting to speculate that the observed shifts in GBP are representative of GALT immune reconstitution since the gut microbiome assists in the maintenance of the gut integrity. Future directions should include gut biopsies to confirm whether the gut flora is associated with immune recovery of the gut, and its correlation with the blood compartment.

It is known that members of *Lactobacillales* aid preserving gut integrity acquiring the name of 'beneficial bacteria'[19]. In the context of HIV, some studies suggest that the intake of probiotics is associated with an increase of systemic CD4 T cells[38,39] and since HIV infection can result in the reduction of *Lactobacillus* species[37], it is not surprising that individuals in Group 1 also

showed the most profound reduction in proportions of *Lactobacillales*. Furthermore, it is suggested that improper balance of Th17/Tregs during HIV infection can result in disease progression[60] and since *Lactobacillales* are capable of regulating different types of immune cells such as Tregs[19], a shift in the proportions of *Lactobacillales* may contribute to HIV pathogenesis.

In contrast to *Lactobacillales*, little is actually known about the relevance of most organisms that live in the gut (despite the fact that they comprise more cells in our body than our own eukaryotic cells) but there is some scarce interesting evidence. Healthy children had increased proportions of *Lactobacillales*, *Enterobacteriales*, *Rhizobiales*, and *Xanthomonadales* when compared to malnourished children[61]. On the other hand, members of *Pseudomonadales*, *Xanthomonadales* and *Lactobacillales* were shown to protect crops from fungal infection[62] and the notion of treating soil with 'beneficial' bacteria in a similar fashion as using probiotics to treat certain conditions has been considered (http://www.secondgenome.com/our-blog/page/7/). Although extrapolating these bacterial associations into an HIV context might be quite difficult, the fact that these similar orders of bacteria were found in significantly less proportions in Group 1 with low CD4% (*Pseudomonadales, Xanthomonadales, Aeromonadales, Rhizobiales* and *Neisseriales*), supports the idea that perhaps they might have a 'beneficial' effect during HIV infection.

This investigation is subjected to several limitations. Our study cohort comprised only 13 participants, of which we only considered 11 individuals at baseline. Consequently, we defined classes using percentages of lymphocytes

rather than absolute counts since in general lymphocyte counts are more volatile than percentages. Therefore, to reduce variability and find significant differences between groups, a bigger sample size would be needed. Moreover, after initiation of therapy, lymphocyte percentages tend to stabilize quicker than absolute counts thereby becoming less informative in terms of immune reconstitution. By the same token, sample size could also explain why sCD14 did not achieve significant levels. However, in the case of LPS, we observed inhibition, which is a common limitation of the assay[63]. Lastly, this investigation cannot determine causality. Nevertheless, our results showed that patient B did not have a significant change in his GBP after HIV infection and based on his GBP at baseline this patient was Group 2, thus we speculate that certain GBP could be advantageous at the time of infection but this needs further investigation.

It is clear that certain compositional and/or functional states of the human gut microbiome are beneficial for the host as they can prevent pathogenic infections and may regulate various immune responses. This may impact the likelihood of microbial translocation[40]. In HIV infection, microbial translocation is predicative of chronic immune activation[32-34], which is one of the strongest predictors of disease progression[64-66] therefore understanding changes of the composition of the gut flora will be important if this information is to be translated to clinical practice. This study identified specific GBP associated with different levels of CD4% in the blood during acute/early HIV infection that could be also

representative of the HIV dynamics in the GI tract and thereby provide a novel and less invasive screening method for gut immune health.

Table 2.1: Differences in clinical variables between groups demarcated by GBP

Comparison between clinical and immunological variables on groups separated by gut bacterial profiles. There was statistical difference between groups in CD4%, CD8% and VL when subjected to a two-tailed Mann-Whitney. Soluble CD14 almost achieve significance levels.

Clinical Variable	Group 1 (median)	Group 2 (median)	*P*-value
EDI[1]	3	3	0.43
CD4%	9.5	33	**0.01***
CD4 Absolute Counts (cells/µL)	398	614	0.16
CD8%	70.5	45	**0.01***
CD8 Absolute Counts (cells/µL)	2612	856	0.23
Viral Load (log_{10} copies/mL)	6.9	4.4	**0.01***
sCD14 (ng/mL)	1458.5	991.6	0.07
LPS (pg/mL)	0.344	0.317	0.79

[1]EDI = Estimated duration of infection
*Denotes level of significance achieved ($P < 0.05$)

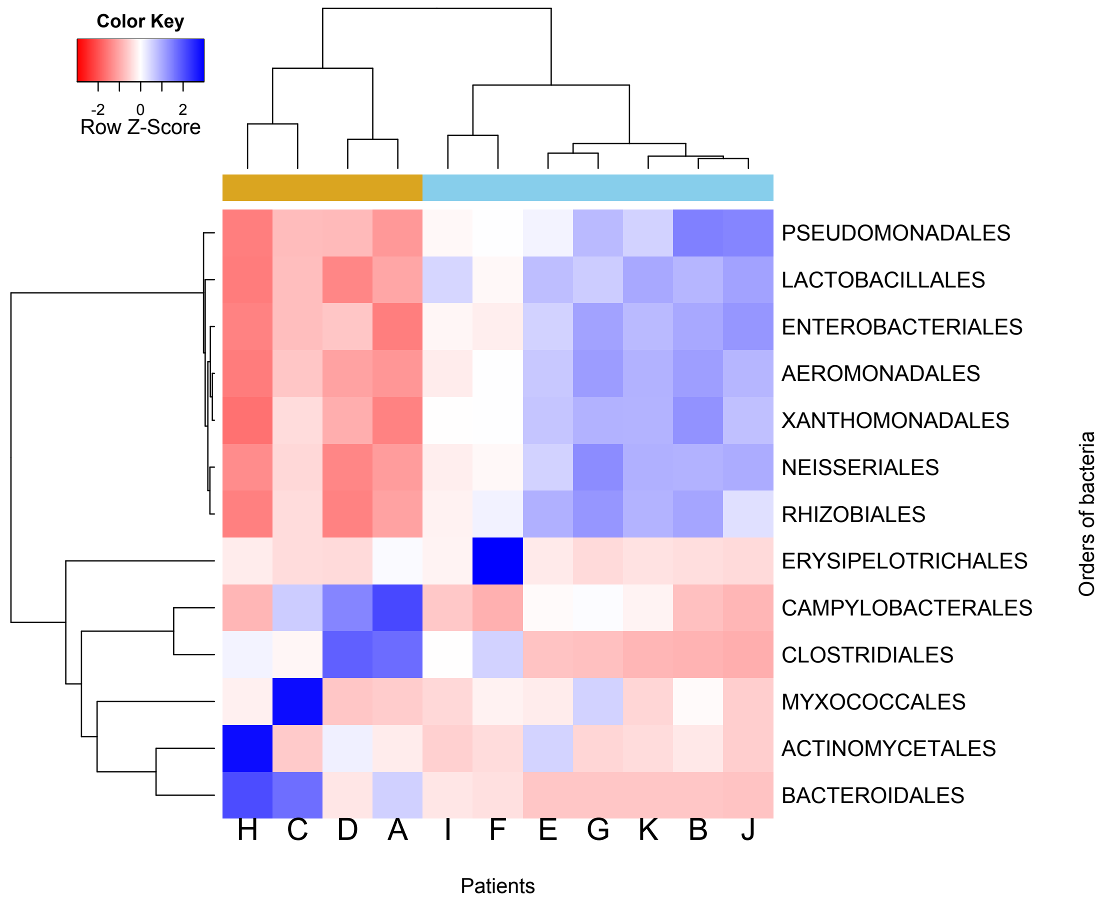

Figure 2.1: Unsupervised clustering at baseline.

Gut bacterial profiles separated our participants in two main groups colored as gold and blue representing participants with low and high CD4% (Group 1 vs. Group 2) respectively. Orders of bacteria are separated into two main groups clustered by correlation. Overall, bacteria in the same cluster are positively correlated whereas bacteria in different clusters are negatively correlated. Blue and red correspond to enrichment or depletion on proportions of bacteria respectively.

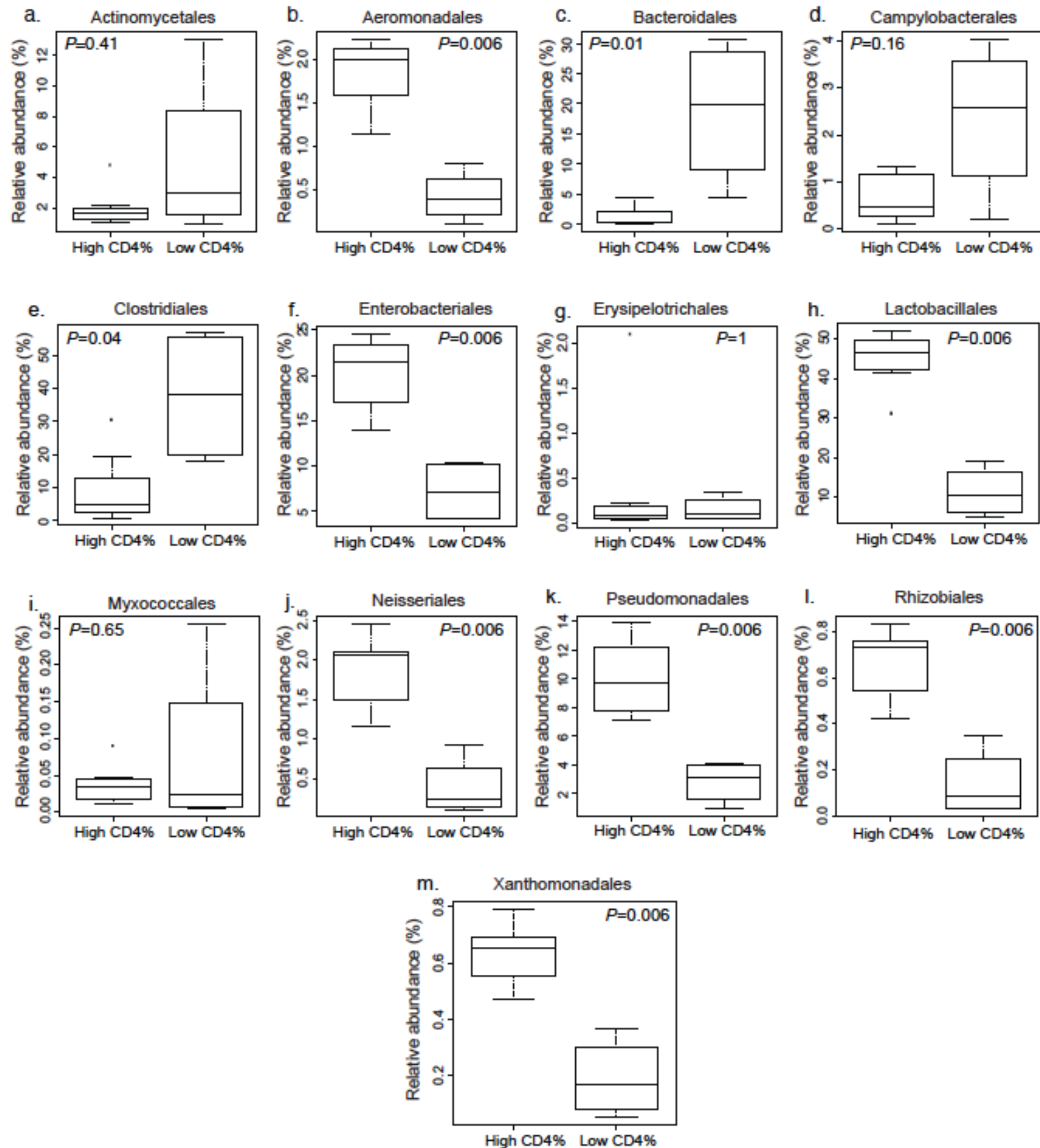

Figure 2.2: Bacterial differences between participants with low vs. high CD4% (Group 1 vs. Group 2) at baseline*.*

Supplementary Figure 1: Comparison of gut bacterial profiles of participant with high vs. low CD4% at baseline and before ART. Participants with low CD4% exhibited significant reduced proportions of *Aeromonadales, Enterobacteriales, Lactobacillales, Neisseriales, Pseudomonadales, Rhizobiales and Xanthomonadales* and significant increase proportions of *Clostridiales* and *Bacteroidales* when they were subjected to two-sided Mann-Whitney test.

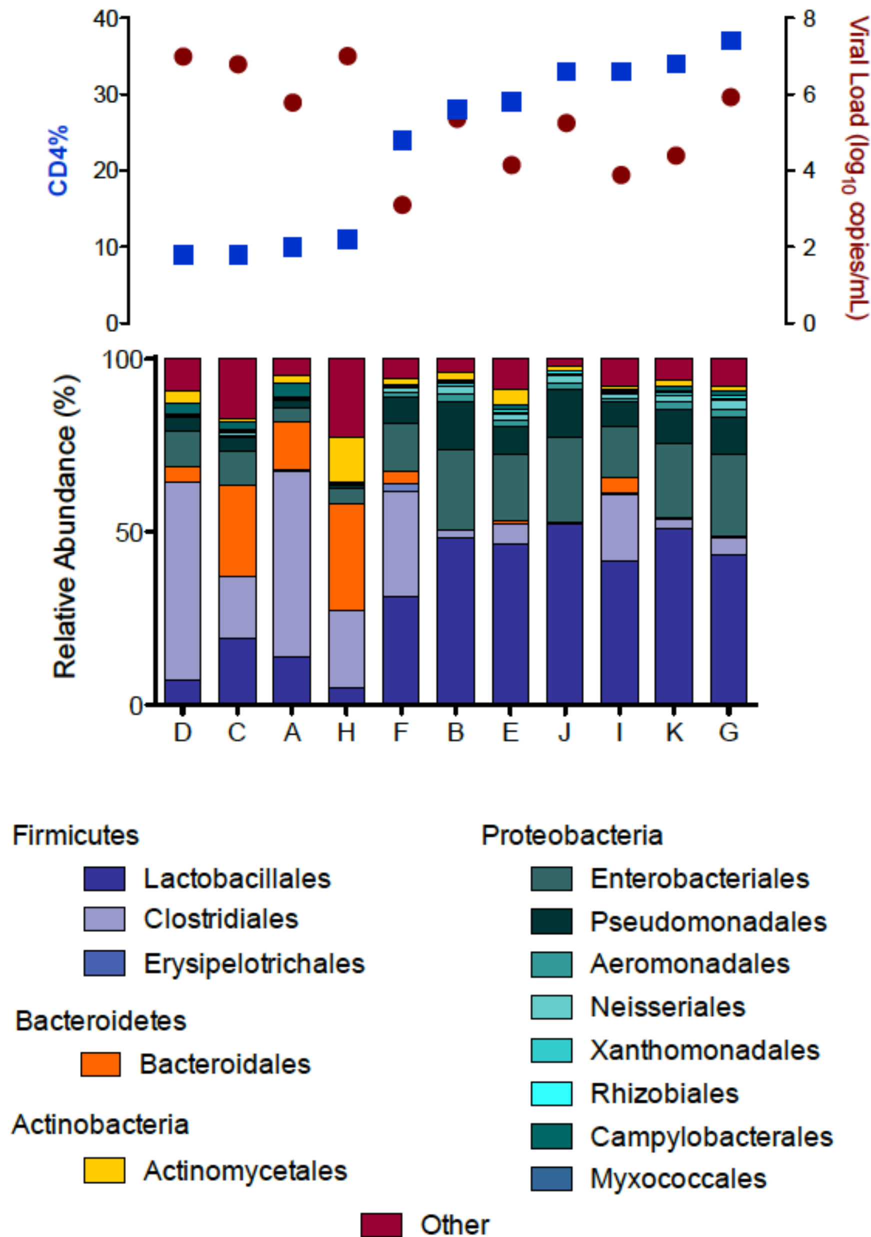

Figure 2.3: Overview of CD4%, VL and GBP at baseline

Participants' CD4% and VL are colored blue and red respectively and their corresponding GBP is in the bottom. Participants with lower CD4% and higher VL exhibit lower proportions of *Lactobacillales*. Orders of bacteria are grouped by phyla and orders in the same phyla share similar colors.

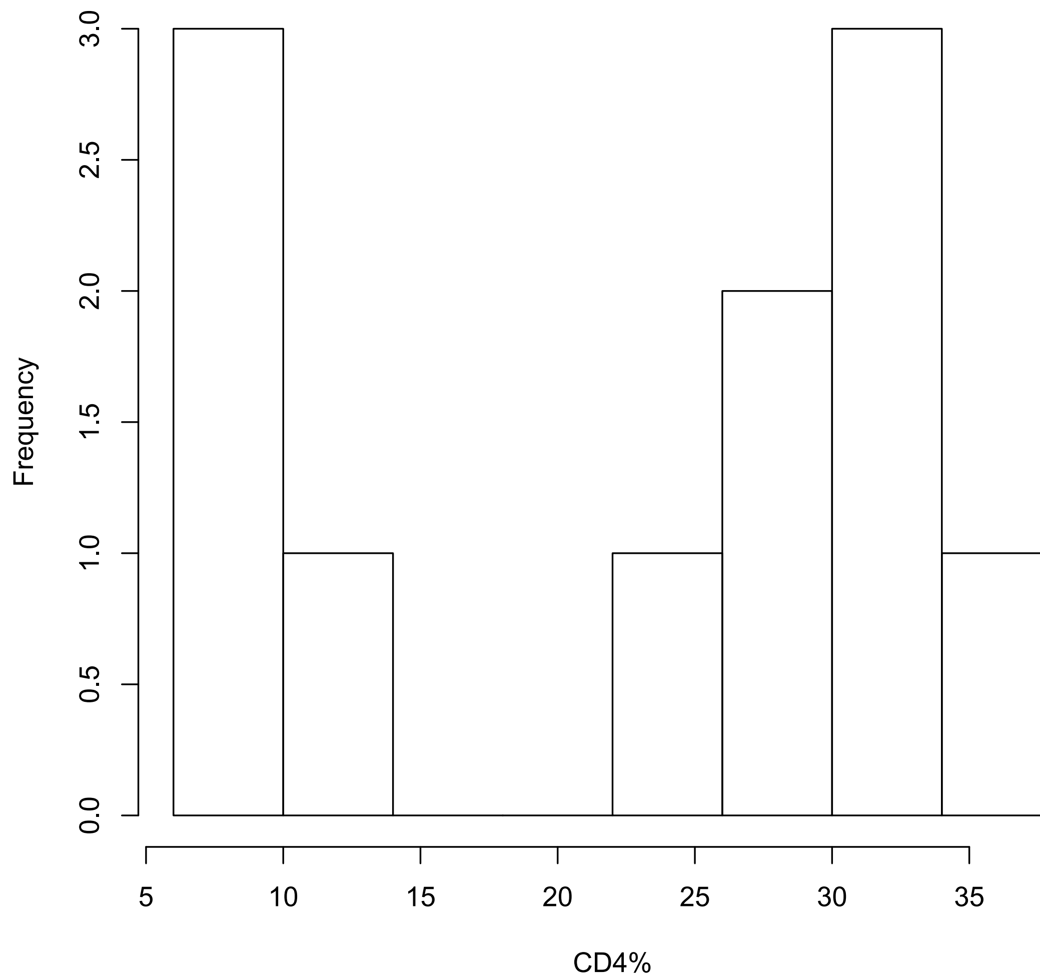

Figure 2.4: Distribution of CD4% at baseline

There is a bimodal distribution of CD4% in acute/early and not-treated HIV infection.

Table 2.2: Supervised classification statistics

At baseline, GBP segregated participants in two groups that were associated with two distinct distributions of CD4% Different CD4% values representing CD4% intermediate values of these two distributions were tested to find the value of CD4% that maximized classification. The maximum accuracy achieved was 0.88 with values of CD4% of 14%-16% (in bold).

CD4%	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Balanced accuracy
14	**0.75**	**1.00**	**1.00**	**0.06**	**0.88**
15	**0.75**	**1.00**	**1.00**	**0.06**	**0.88**
16	**0.75**	**1.00**	**1.00**	**0.06**	**0.88**
17	0.76	0.50	0.96	0.11	0.63
18	0.75	0.40	0.94	0.11	0.58
19	0.75	0.33	0.92	0.11	0.54
20	0.75	0.33	0.92	0.11	0.54
21	0.75	0.33	0.92	0.11	0.54
22	0.77	0.44	0.90	0.22	0.61
23	0.76	0.36	0.87	0.22	0.56

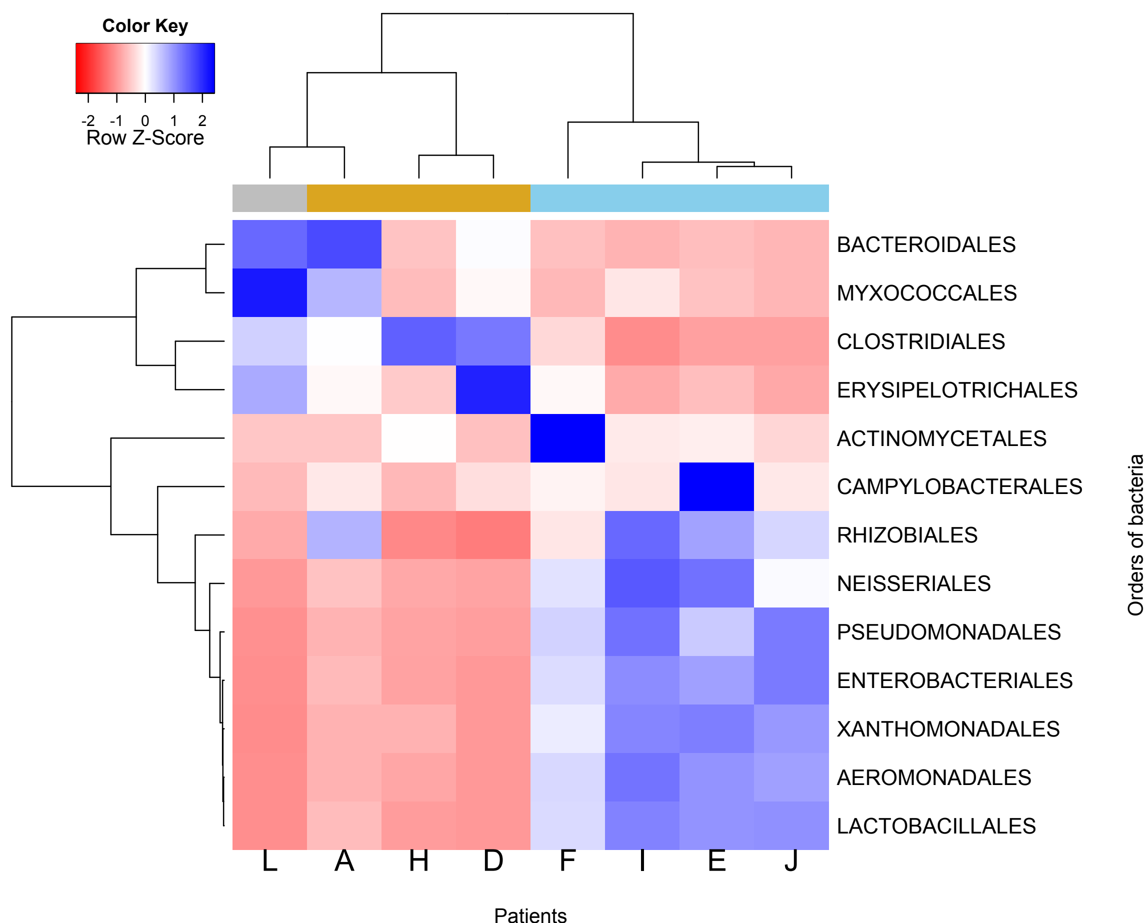

Figure 2.5: Unsupervised clustering at week 24 of ART

Comparisons between gut bacterial profiles after the initiation of antiretroviral therapy. Gut bacterial profiles separated our participants in the two main groups as baseline Participants colored as golden have lower CD4% compared to participants colored as blue and have same classification as baseline. Participant in gray was discarded from baseline analysis because of antibiotic use.

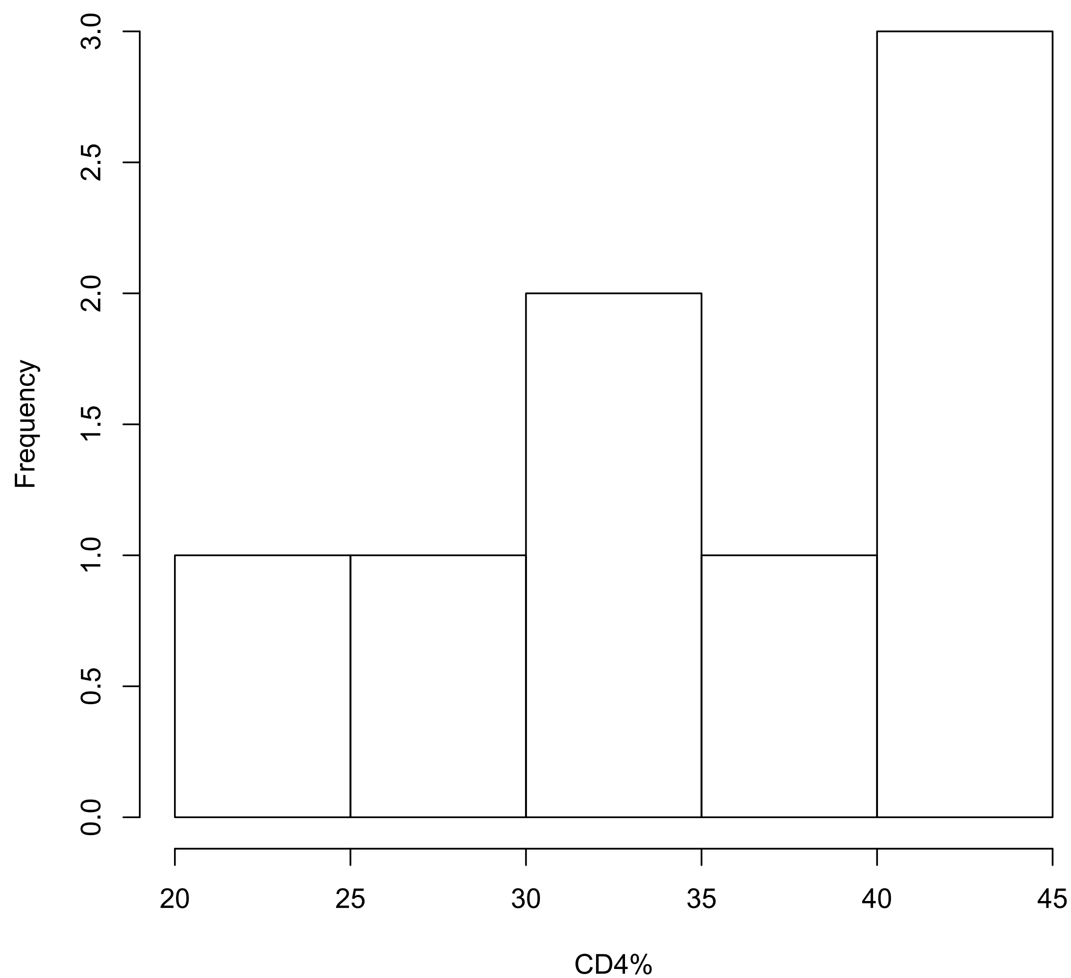

Figure 2.6: Distribution of CD4% at week 24 of ART

There is a unimodal distribution of CD4% after 24 weeks of ART.

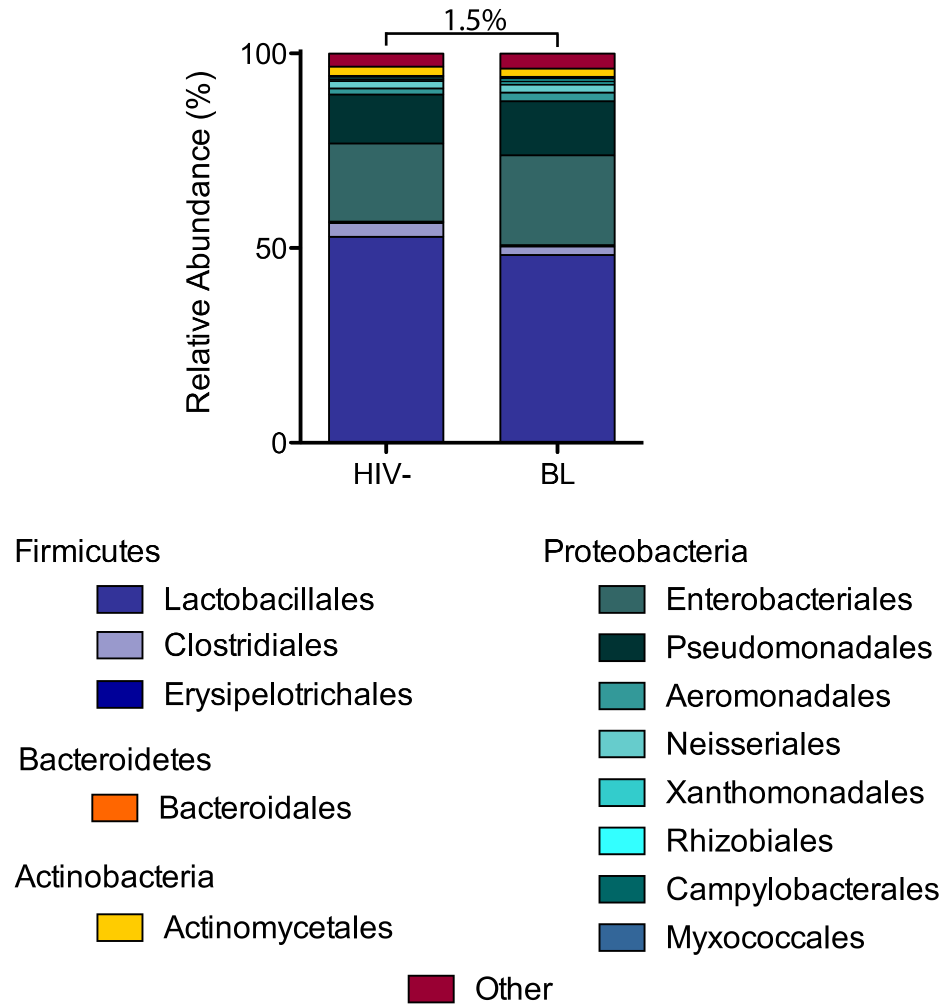

Figure 2.7: Changes in gut flora before and after HIV infection

Gut bacterial profiles of patient B taken before and after HIV infection showed a 1.5% difference. HIV did not impact significantly the composition of the gut flora. Orders of bacteria are grouped by phyla and orders in the same phyla share similar colors.

ACKNOWLEDGEMENTS

Chapter 2, in part, has been included in the submission for publication. Josué Pérez-Santiago, Sara Gianella, Marta Massanella, Celsa A. Spina, Maile Y. Karris, Susanna R. Var, Derek Patel, Parris S. Jordan, Jason A. Young, Susan J. Little, Douglas D. Richman and Davey M. Smith. Gut *Lactobacillales* are Associated with Higher CD4 and Less Microbial Translocation during HIV Infection. The dissertation author was the primary investigator and author of this material.

Chapter 3

The role of *Lactobacillales* in HIV pathogenesis

Abstract

Early HIV infection is characterized by a dramatic depletion of CD4 T-cells, especially in the gastrointestinal tract, and is associated with translocation of bacterial products into the blood. This study identified that higher proportions of *Lactobacillales* in the distal gut of recently HIV-infected individuals were associated with lower markers of microbial translocation, higher CD4% and lower viral loads before antiretroviral therapy was started. During antiretroviral therapy, higher proportions of gut *Lactobacillales* were associated with higher CD4%, less microbial translocation, less systemic immune activation, less gut T lymphocyte proliferation and higher CD4% in the gut. Shaping the gut microbiome, especially proportions of *Lactobacillales*, could help to preserve immune function during HIV infection.

Introduction

The gut associated lymphoid tissue (GALT) is a crucial part of the immunological network that maintains the integrity of the gastrointestinal (GI) tract against gut microbes[67]. Early human immunodeficiency virus (HIV) infection results in the substantial depletion of CD4 T-cells, preferentially in the GALT[25-27]. This lymphocyte loss promotes mucosal immune dysfunction, increased permeability of the gut and ultimately translocation of bacterial products[31], which contribute to chronic immune activation in the setting of HIV infection[32-34] and is one of the strongest predictors of HIV disease progression[64-66]. Antiretroviral therapy (ART) seems to at last partially restore gut integrity[36,68] with marginal reduction in microbial translocation, but not to the levels seen in HIV uninfected persons[32,34].

The GI tract is colonized by numerous commensal microorganisms, which can be identified by next generation sequencing. Next generation sequencing of the gut microbiome has revealed that certain disease states, such as inflammatory bowel disease, obesity, GI surgery and diabetes may be associated with modified gut flora[20-22]. Certain microbes may interact with the GALT to preserve gut integrity[69], thereby decreasing the likelihood of translocation of microbial products[40]. For example, studies have implicated that consumption of probiotics and specifically *Lactobacillales* could modulate inflammatory responses, eradicate potential pathogens, and reduce gut permeability[40,70-74]. Manipulation of the gut flora may therefore benefit immune recovery during HIV

infection. In this study, we longitudinally characterized the changes of *Lactobacilalles* during acute and early HIV infection and to examine the effects of ART by associating clinical and immunological factors before starting and during ART.

Methods

Gut bacterial profiles and clinical variables and microbial translocation markers

Classification of bacteria, peripheral lymphocyte profiles, viral load (VL) and markers of microbial translocation, i.e. lipopolysaccharide (LPS) and Soluble CD14 (sCD14), described in previous chapters were used. Briefly, all participants started ART within a week of study enrollment and were followed for 48 weeks with specimens collected approximately every four weeks. Gut bacterial profiles (GBP) from participants L and M were excluded from baseline analysis because of antibiotics use.

Nucleic acid extraction and quantification from peripheral blood mononuclear cells (PBMC) and semen

Genomic DNA was extracted from 5 million PBMC for each time-point using QIAamp DNA Mini Kit (Qiagen, CA) per manufacturer's protocol. Extracted DNA was eluted in 100µl elution buffer and total proviral HIV-1 DNA was quantified by real-time PCR in an ABI 7900HT thermocycler (Applied Biosystems, CA) with 0.005 M ROX as passive reference. Amplification was performed in a 50 l reaction. Master mix consisted of 10 l DNA extract, TaqMan Environmental Mastermix 2.0 (Applied Biosystems, CA), virus specific PCR primers mf299 and mf302 (1µM each) and two DNA locked nucleic acids (LNA) detection probes Ri15 and Ri16 (0.1µM each) as previously described[75]. The reaction used the following cycling conditions (1) 2 minutes at 50°C, (2)

initial activation at 95°C for 10 minutes and (3) 60 cycles of denaturation at 95°C for 15 seconds followed by 60 seconds of annealing/extension at 60°C. Cellular input was normalized with beta-actin PCR using the primers Beta-Actin-959F: 5'-CTGGCACCCAGCACAATG-3' and Beta-Actin-1026R: 5'-GCCGATCCACACGGAGTACT-3', and detection probe Beta-Actin-980T: 5'-f-TCAAGATCATTGCTCCTCCTGAGCGC-q-3' and expressed in HIV-1 DNA copies numbers/1 million actin cells equivalents.

DNA extraction from seminal plasma and quantification of cytomegalovirus (CMV) in seminal plasma and stool DNA was done as previously published[76].

Flow cytometry

Briefly, fresh whole blood was collected from participants at weeks 0,12, 24, and 48 of study and processed using density gradient centrifugation to obtain PBMCs. The PBMCs were washed and aliquoted into tubes at a concentration of 1 million PBMCs/200uL. Cells were incubated with antibodies for surface marker staining, before fixation and permeabilization for intracellular assays (eBioscience, San Diego, CA). The following antibody combinations were used to evaluate immune activation and proliferation in T-cells: i) HLA-DR–FITC, CD45RO–PE, CD38–PE-Cy7, CD27–APC, CD3–APC-Cy7, CD4–PerCP-Cy5.5 and CD8–Pacific Blue, and ii) Ki67–FITC, CD45RO–PE, CD27–APC, CD3–APC-Cy7, CD4–PerCP-Cy5.5 and CD8–Pacific Blue, Becton Dickinson and Company (Franklin Lakes, New Jersey). These measures were used to also assess naïve ($CD45RO^-CD27^+$), central memory ($CD45RO^+CD27^+$) and effector memory

($CD45RO^{+}CD27^{-}$) in CD4 and CD8 T-cell subsets. Samples were run on the BD FACSCanto (BD Biosciences, San Jose, CA) and data analyzed with FlowJo software (Tree Star Inc, Ashland, OR).

Flow cytometry of gut biopsies was performed on viably thawed mucosal mononuclear cells. Proliferation was assayed using the following combination of antibodies: Ki67–FITC, CD45RO–PE, CD27–APC, CD3–APC-Cy7, CD4–PerCP-Cy5.5, CD8–Pacific Blue, and Aqua-L-D–Amcyan. Samples were run on the BD FACSCanto (BD Biosciences, San Jose, CA) and data analyzed with FlowJo software (Tree Star Inc, Ashland, OR).

Regression analyses

Regression analyses were performed using R statistical software. Cross-sectional associations between GBP and clinical and immunologic variables were evaluated using fixed effects linear models. For longitudinal analyses, mixed effects linear models were utilized for analysis of longitudinal data by adjusting for repeated measurements. This was accomplished using the package *lme4* in R. To evaluate significance of mixed models, p-values were calculated using a Markov-Chain Monte Carlo sampling technique included in the package *languageR*. Normality of each order of bacteria and clinical variables was assessed using a Shapiro test with a significance cut-off of $p < 0.05$ indicative of non-normalcy.

Results

Immune and clinical correlates of *Lactobacillales* before ART

Before initiating ART, proportions of gut *Lactobacillales* were significantly correlated with CD4% (P=2.8x10^{-5}), CD4/CD8 T-cell ratio (P=0.0003) and CD4 counts (P=0.03), and negatively associated with CD8% (P=0.002), but only a negative trend with CD8 counts (P=0.09, **Figure 3.1*a-e***). Proportions of *Lactobacillales* in gut were also negatively correlated with VL (P=0.03) and sCD14 (P=0.04, **Figure 3.1*f-g***), but there was no association with LPS (P=0.88, **Figure 3.2*h***). Despite an association with a marker of microbial translocation (sCD14), there was no observed relationship between *Lactobacillales* and either CD4 and CD8 lymphocyte activation and proliferation (measured as HLA-DR^+ or $CD38^+$ T cells for activation and $Ki67^+$ for proliferation) or gut CD4 lymphocyte proliferation (percentage of $Ki67^+$ of CD4 T cells, data not shown). There was, however, a negative association between higher proportion of gut *Lactobacillales* and lower gut CD8 lymphocyte proliferation in the central memory subset (percentage of $Ki67^+$ of $CD8^+CD45RO^+CD27^+$, P=0.04, **Figure 3.1*i***).

Since duration of HIV infection is associated with immune activation[77] and participants had variable EDI at baseline, we evaluated if EDI at baseline influenced the associations between baseline proportions of *Lactobacillales* and lymphocyte activation and proliferation of lymphocytes in the periphery. As might be expected, participants with a more recent EDI (≤4 weeks or acute phase of the infection, n=8) at baseline showed a higher T-cell activation and proliferation

than participants with longer EDI at baseline (>4 weeks, n=3, **Figure 3.2)**; however, baseline proportions of *Lactobacillales* were not associated with EDI (data not shown). Interestingly, patient H had extremely high levels of immune activation driving all regressions and we were unable to find any clinical explanation for this, although he was the only one with high proportions of *Actinomycetales* (~13%) compared to less than 5% for everyone else.

Immune and clinical correlates of *Lactobacillales* during ART

To determine the impact of ART on the associations between proportions of *Lactobacillales*, clinical variables (VL, CD4%, CD4 counts, CD8% and CD8 counts), markers of translocation (sCD14 and LPS), and lymphocyte activation (HLA-DR, CD38) and proliferation (Ki67) markers, we performed regression analyses for all variables at each study time point. As expected, ART was very effective at suppressing VL, and only one participant (patient H) had detectable levels of VL at weeks 24 and 48 (2.70 and 1.91 HIV RNA $\log_{10}$ copies/mL, respectively). Similarly, all participants increased their CD4% during ART (median +9%, range 4-25%, *P*=0.001, **Figure 3.3*a*)**, and there was a trend for gut proportions of *Lactobacillales* to be positively associated with CD4% gains (*P*=0.07, **Figure 3.3*a***). The relationships between *Lactobacillales* and CD4%, CD4/CD8 ratio and CD8% remained consistent with baseline results at week 24 (*P*=0.01, *P*=0.01, *P*=0.05, respectively), but not at week 48 (**Figure 3.4*a-c***); however, there were no associations between proportions of *Lactobacillales* and CD4 or CD8 T-cell counts at either week 24 or 48 (data not shown). After 48

weeks of ART, increased proportions of *Lactobacillales* were positively correlated with increased CD4% in the gut (*P*=0.04, **Figure 3.4*d***).

Evaluation of lymphocyte activation demonstrated a strong negative correlation between the proportions of *Lactobacillales* and CD8 T-cell activation in the blood after 24 weeks of ART ($CD45RO^+HLA\text{-}DR^+$, *P*=0.002, $CD45RO^+CD38^+$, *P*=0.01, **Figure 3.4*e-f***). After 48 weeks of ART, the proportions of *Lactobacillales* were still negatively associated with $CD38^+$ lymphocyte activation (*P*=0.04), but only a trend remained for $HLA\text{-}DR^+$ (*P*=0.09) in CD8 T cells (**Figure 3.4*e-f***). The proliferation of central memory CD4 T cells in the gut was negatively associated with *Lactobacillales* at week 24 of ART (percentage of $Ki67^+$ of $CD4^+CD45RO^+CD27^+$, *P*=0.05), but not after 48 weeks (*P*=0.22**, Figure 3.4*g***). Regarding microbial translocation, the negative association between sCD14 and *Lactobacillales* observed at baseline was lost at week 24 but regained after 48 weeks of ART (*P*=3.7×10^{-4}, **Figure 3.4*h***). Throughout the study, sCD14 was inversely correlated with *Lactobacillales* (*P*=0.02) and time on ART (*P*=0.04, **Figure 3.3*b***). In contrast, LPS showed an isolated negative association at week 24 (*P*=0.03, **Figure 3*i***).

HIV latent reservoir and CMV shedding

Since changes in immune activation may impact the HIV latent reservoir, we also evaluated HIV DNA levels in PBMCs, but we did not observe any associations between HIV proviral DNA and *Lactobacillales* after 24 or 48 weeks of ART. Proviral DNA did not correlate with CD8 T-cell activation (**Figure 3.5*a-b***).

There was a strong negative correlation, however between proviral DNA levels and activation of CD4 lymphocytes ($CD45RO^{+}HLA\text{-}DR^{+}$, $P=0.007$) at week 24 of ART, but not at week 48 (**Figure 3.5*c-d***). We also investigated the relationship between CMV shedding in the gut, the genital tract and *Lactobacillales*, since the presence and magnitude of CMV shedding may influence immune activation[76] and all participants were CMV antibody positive. Only two patients (D and H) had detectable levels of CMV DNA in rectal swabs. Patient D had 2.41 and 2.14 $\log_{10}$ copies per swab at weeks 16 and 24 and patient H had 1.46 $\log_{10}$ copies per swab at week 4. Due to limited sample availability, only nine semen samples were screened for CMV shedding and Patient D exhibited high levels of CMV copies (4.49 $\log_{10}$ copies per swab). The eight remaining samples showed no evidence of CMV, though no samples were available for Patient H. Interestingly, both patients (D and H) exhibited the lowest overall proportion of *Lactobacillales* (<10%) throughout infection and Patient D in particular, had the lowest CD4% overall after 48 weeks of ART.

Discussion

This is the first study to identify associations between specific GBP, higher CD4 counts, CD4%, VL, immune activation, CD4 T-cell proliferation in the gut and evidence of microbial translocation in untreated HIV infection. All of these factors have been previously associated with better HIV disease outcomes[78,79]. The associations between GBP, CD4%, immune activation, and markers of bacterial translocation continued, albeit weakly, during ART that suppressed VL. The important caveat of this study is that the associations between changes in the GBP and the HIV disease markers cannot determine causality.

The human gut prevents the translocation of commensal bacteria through physical barriers (e.g. epithelial tight junctions), biochemical agents (e.g. antibacterial peptides and mucus), and immune mechanisms (e.g. secretory IgA and Toll-like receptor mediated sensing, oxidative bursts)[67,80-83] During early HIV infection, gut populations of *Bifidobacteria* and *Lactobacillus* species are reduced[37], the GALT is depleted[34], the gut barrier is compromised and translocation of bacterial products can occur. The translocation of these products is associated with HIV disease progression, most likely through persistent immune activation[32,34]. This study aimed to further evaluate these connections by investigating 13 men who were recently infected with HIV and started ART within 14 weeks of their infection. By evaluating multiple factors observed during acute and early HIV infection and by treating the HIV infection at the earliest stages possible, this study aimed to identify the correlates associated with

optimal immune recovery and preservation. The current study extends previous observations by linking the constitution of the gut microbiome itself to immunologic and virologic dynamics during recent HIV infection and during ART, specifically identifying that higher proportions of *Lactobacillales* in the gut are associated with markers that are predicative of better HIV outcomes including higher CD4 percentages, lower VL, and less evidence of microbial translocation. Since *Lactobacillales* can modulate anti-inflammatory responses and immune cells (i.e. Tregs), improve gut integrity and reduce gut permeability in other conditions[19,70-74], it is interesting that this bacterial order would be identified as the main component of GBP associated with markers of better HIV outcomes.

There are a number of limitations that should be considered. First, this investigation was conducted in the setting of a randomized double-blind controlled trial of maraviroc versus placebo combined with standard ART (atazanavir boosted with ritonavir and tenofovir plus emcitritabine), but the overall study remains blinded to maraviroc use because participant enrollment continues. Since the associations between GBP and clinical and immunological variables identified at baseline and before the start of ART for all participants, largely held throughout follow-up, we do not think that maraviroc use contributed significantly to the observations, and that delaying these results until the unblinding of the study are not necessary. Second, although next generation sequencing allowed us to conduct large-scale metagenomics analyses of the distal gut microbiome, the analysis is limited by the classification of sequences. To focus our study on the main drivers of gut bacteria across all participants, we

only considered bacterial populations shared across individuals. Bacterial sequences that were not identified in all participants were classified as “Other”, and these “Other” populations may be very informative and should be investigated further. Along these lines, the study only considered classification at the order level and a finer grained classification may provide additional insight. To this end, all bacterial sequences have been publically deposited for study by outside groups. Third, the study evaluated both sCD14 and LPS as measures of microbial translocation, and as reported previously[34,84] sCD14 showed the largest associations. The lack of association with LPS could be because although microbial translocation occurs in early phases of infection, levels of LPS do not increase until later HIV stages, while levels of sCD14 increase earlier[34]. Fourth, in a very small number of samples, this study investigated if CMV shedding in the distal GI Tract or semen demonstrated similar associations to classified GBP. Although, the associations between rectal CMV shedding and lower CD4 T-cell count and percent were provocative among two of the participants, the numbers are too small to draw conclusions but should provide testable hypotheses for new investigations.

It is increasingly evident that the human immune system is linked to the GI system[19], but translating this information to clinical care is lacking, especially for HIV infection. While the gut microbiome is influenced by birth delivery methods, host genetics, household contacts, age, geography, etc.[11], it is unknown if the microbiome can be shaped with agents like prebiotics, probiotics or targeted antibiotics for beneficial outcomes. This study identified that increasing

Lactobacillales in the gut could be important for recovering and preserving immune system function during HIV infection, and a promising clinical target may be HIV infected individuals who are able to suppress their VL with ART but unable to sufficiently recover their CD4 T-cell counts.

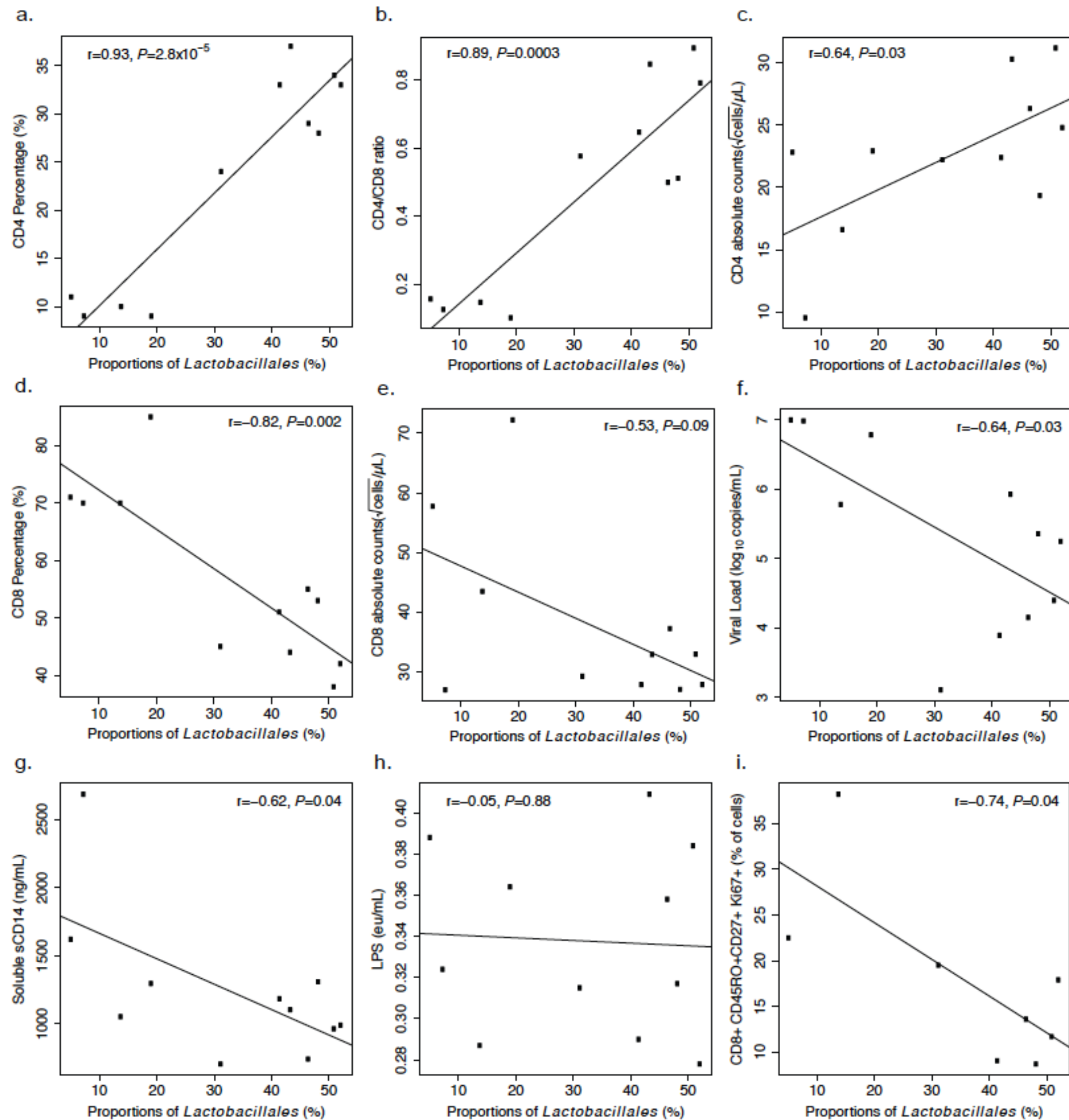

Figure 3.1: Associations of *Lactobacillales* with clinical and immunological variables before ART

There is a positive association with (A) CD4%, (B) CD4/CD8 ratio, (C) CD4 counts and a negative correlation (or trend) with (D) CD8%, (E) CD8 counts, (F) Viral Load, (G) soluble CD14, (H) LPS, and (I) Gut CD4+ T-cell proliferation. All these correlations suggest that higher proportions of Lactobacillales are beneficial for the host in the absence of ART.

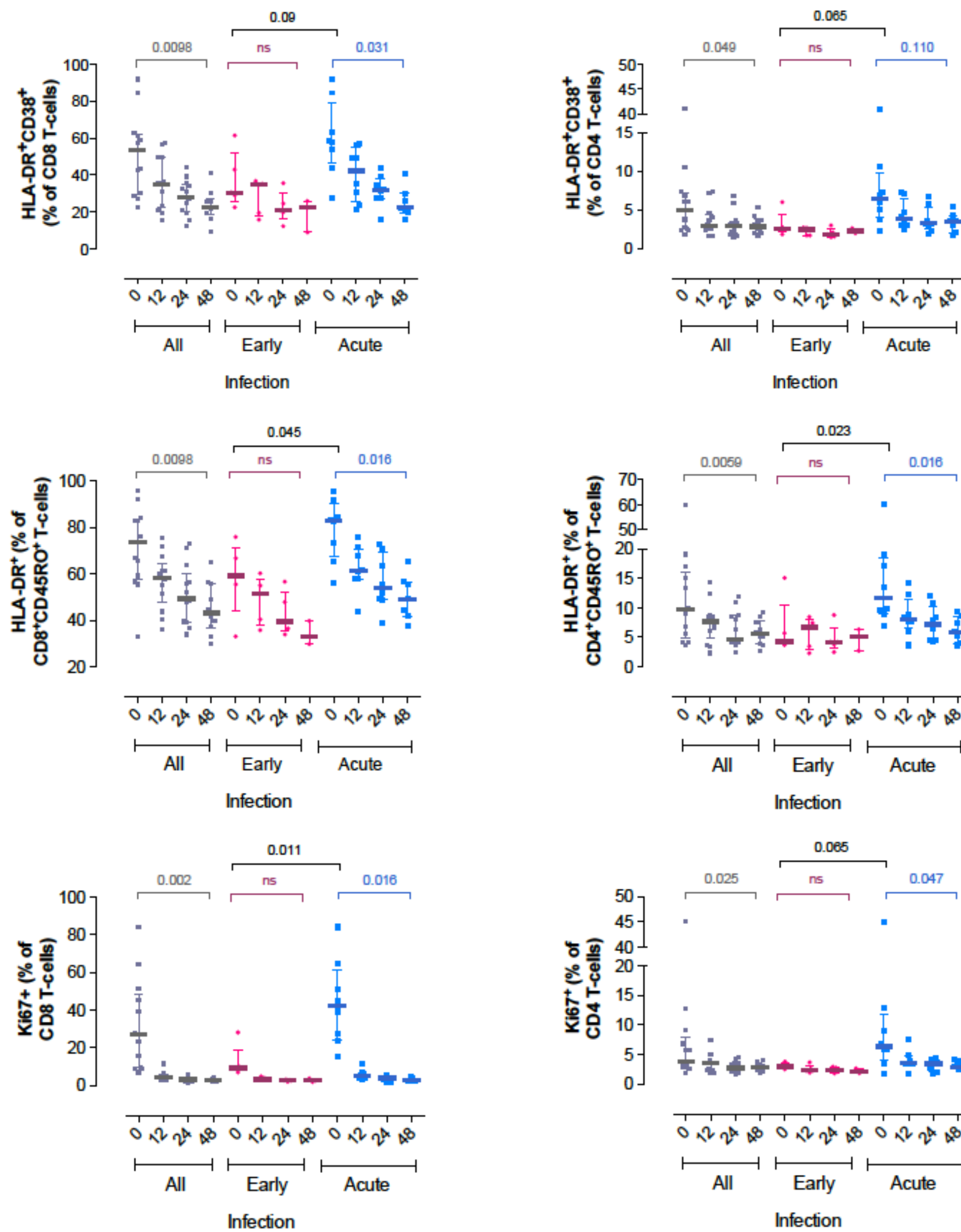

Figure 3.2: Immune activation and proliferation at baseline between early vs. acute infected individuals

Phenotypic analysis of T lymphocytes subsets in all participants and divided by early vs. acute phase of infection. Participants more recently infected exhibited higher immune activation and proliferation than participants in the early phase.

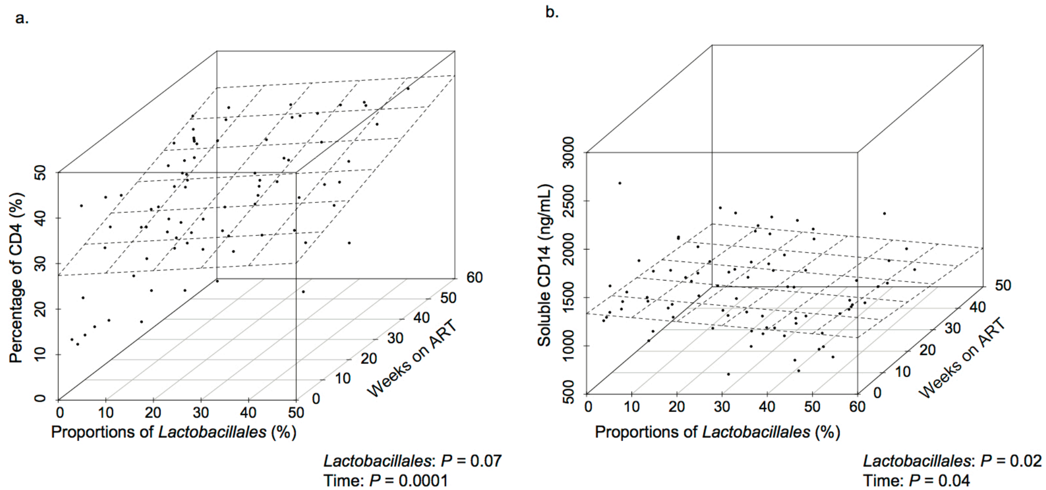

Figure 3.3: Longitudinal associations of *Lactobacillales* with CD4% and soluble CD14

(A) CD4% increases with the initiation of ART (P=0.001) and there is a positive trend between CD4% (P=0.07) and *Lactobacillales*. (B) Less microbial translocation is associated with time on therapy (P=0.04) and higher proportions of *Lactobacillales* (P=0.02).

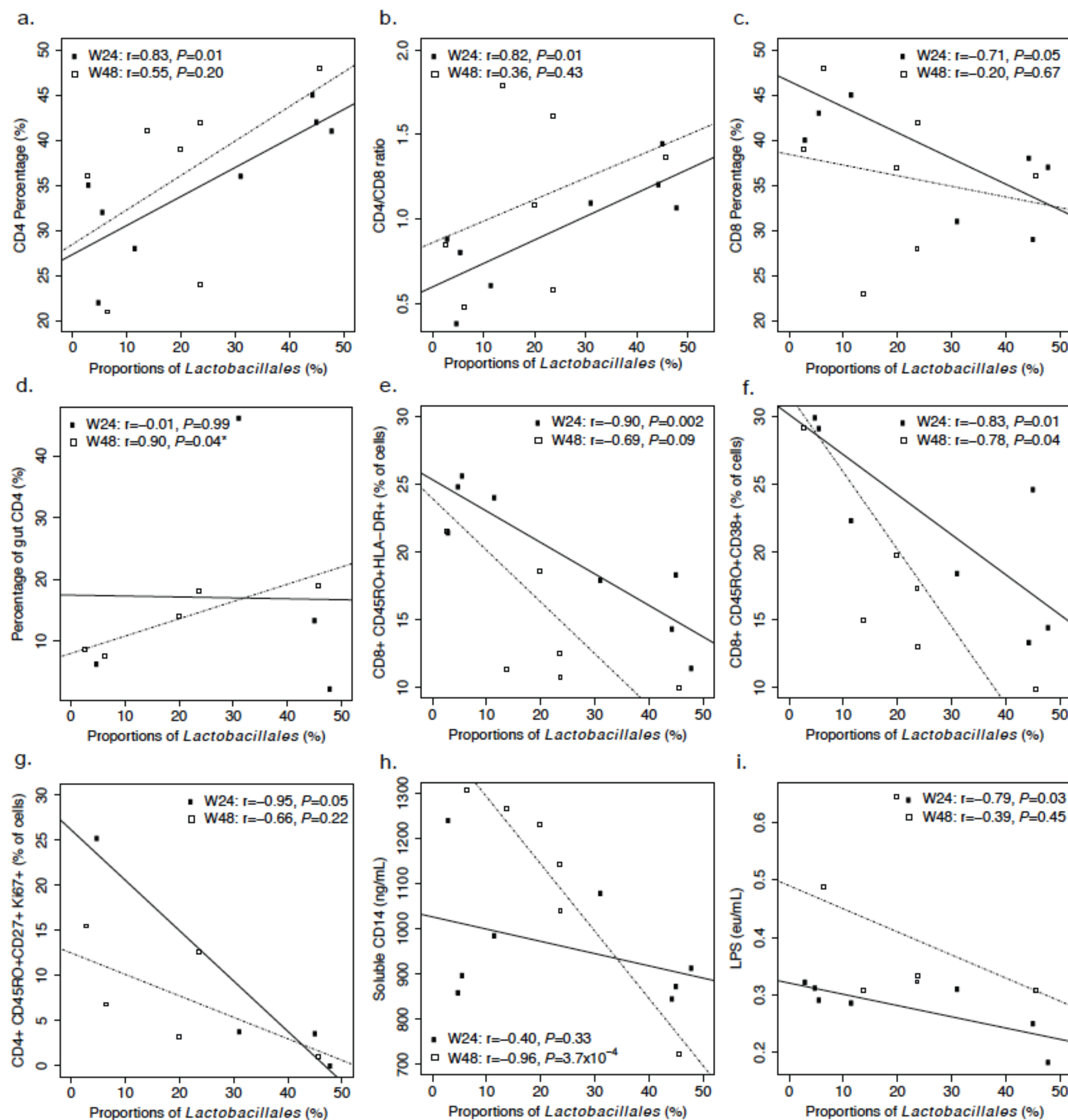

Figure 3.4: Associations of Lactobacillales with clinical and immunological factors during ART.

Week 24 is represented by solid dots and lines whereas week 48 by empty dots and dotted lines. Consistently with baseline correlations, at week 24 there is a positive association with (A) CD4%, (B) CD4/CD8 ratio and a negative correlation (C) CD8%. Additionally, higher gut *Lactobacillales* is associated with higher CD4% in the gut (D), less CD8+ T-cell activation (E, F), less CD4+ T-cell proliferation (G), CD4% in the gut (H), and less microbial translocation (I, J). Associations were independent of ART and suggest that higher proportions of *Lactobacillales* are associated with better immune health.

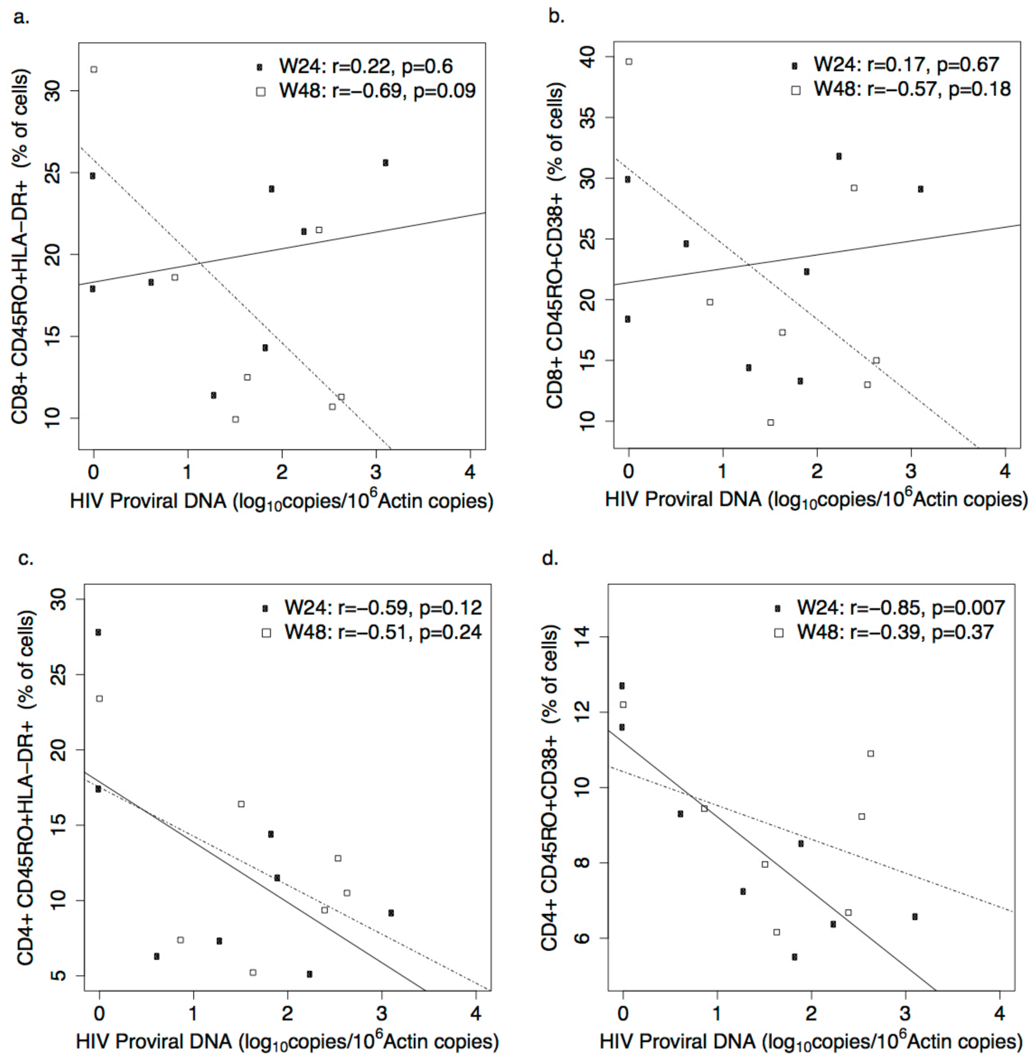

Figure 3.5: Association of HIV proviral DNA and activation markers

Cross-sectional associations between HIV proviral DNA and T-lymphocyte immune activation. (A and B) Solid and dotted lines represent week 24 and 48 respectively. There was no association between HIV proviral DNA and CD8 lymphocyte activation at weeks 24 and 48. (C) There was no association in CD4 lymphocyte activation (HLA-DR^{+}) but (D) there was a strong negative association with CD38^{+} at week 24 but no at week 48.

ACKNOWLEDGEMENTS

Chapter 3 in full, has been included in the submission for publication. Josué Pérez-Santiago, Sara Gianella, Marta Massanella, Celsa A. Spina, Maile Y. Karris, Susanna R. Var, Derek Patel, Parris S. Jordan, Jason A. Young, Susan J. Little, Douglas D. Richman and Davey M. Smith. Gut *Lactobacillales* are Associated with Higher CD4 and Less Microbial Translocation during HIV Infection. The dissertation author was the primary investigator and author of this material.

Future Directions

In this thesis, we conducted a metagenomics analyses to longitudinally characterize the changes of the gut flora during acute/early HIV infection before and after antiretroviral therapy.

First, this study found that the bacterial order of *Lactobacillales* was consistently associated with most clinical and immunological variables that were measured. Other orders of bacteria also showed some associations but with the limited sample size, solid conclusions could not be derived. Hence, further investigation with a larger sample size will be needed to further evaluate these other bacterial orders. Additionally, to complete the mechanistic links between HIV, the human gut and microbial translocation, future evaluations should characterize bacterial populations in blood by next generation sequencing with the goal of determining whether there is selection of translocating bacteria driving systemic immune activation.

Second, for one participant we observed that HIV infection did not impact significantly the composition of his gut flora. Future studies should longitudinally evaluate uninfected individuals who are at high risk of HIV infection in order to characterize the impact of HIV infection on the composition of their gut flora. Thereby, studying the changes of the gut flora in untreated HIV infected and healthy individuals over time and their association with the similar clinical variables will provide a better insight of the overall dynamics of the gut flora and the immune system.

Third, lipopolysaccharide (LPS) and soluble CD14 (sCD14) were used as markers of microbial translocation and we found that *Lactobacillales* were negatively correlated with both markers. Nonetheless, sCD14 showed more consistent associations throughout the study period, highlighting the lack of reproducibility of the LPS detection, which has been reported previously[63]. However, sCD14 is a marker of macrophage activation and is not specific for bacterial translocation. For example, HIV triggers an interferon response, which might stimulate macrophage activation and therefore solubilization of CD14 receptor. An alternative measure that could be useful in future studies would be the quantification of 16s rDNA in blood, which might provide a better estimate of systemic levels of bacterial DNA.

Finally, in this study we observed that low proportions of *Lactobacillales* in the gut of HIV infected individuals were associated with high immune activation and low CD4%. This raises an interesting question: what is the mechanism of CD4 T-cell depletion? Since, HIV infection is associated with high levels of activation and depletion of CD4 T cells, this might perturb the dynamics of the gut flora thereby decreasing the percentage of *Lactobacillales* in the gut. Alternatively, the attempt of the immune system in maintaining gut homeostasis could result in T lymphocyte activation, which then could predispose the CD4 T cells to HIV infection and death. Therefore, hypothetically having high proportions of *Lactobacillales* before HIV infection may reduce localized gut immune activation, thereby reducing the likelihood of T-cell infection and apoptosis. However, it remains unclear if these relationships are also observed in healthy

individuals and should be considered for future work. For HIV infected individuals, a clinical trial testing the effects of prebiotics, probiotics and symbiotics designed to increase levels of gut *Lactobacillales* could address the relationship of the dynamics of HIV infection and the gut flora.

Glossary

1. Bacteroicins – toxin produced by certain bacteria that can be detrimental to closely related bacteria.
2. CCR5 – chemokine (C-C motif) receptor type 5, needed for HIV entry.
3. Prebiotics – nutrients that promote the growth of 'beneficial' microbes.
4. Probiotics – live organisms considered to have a 'beneficial' effect on the host.
5. Symbiotics – combination of prebiotics and probiotics.
6. Th17 or T helpers 17 cells – subset of T lymphocytes that produces interleukin 17.
7. Tregs or regulatory T cells – subset of T lymphocytes that regulates immune responses to infections.
8. Xenobiotics – compounds found in a organisms that are not normally present in the organism. Examples include drugs or toxins.

References

1. Cojocaru, G.S., Rechavi, G. & Kaminski, N. The use of microarrays in medicine. *Isr Med Assoc J* **3**, 292-296 (2001).

2. Weeraratna, A.T., Nagel, J.E., de Mello-Coelho, V. & Taub, D.D. Gene expression profiling: from microarrays to medicine. *J Clin Immunol* **24**, 213-224 (2004).

3. Woelk, C. & Burczynski, M. The clinical relevance of gene expression profiles in peripheral blood mononuclear cells. (2008).

4. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* **389**, 1017-1031 (2007).

5. Bantscheff, M., Lemeer, S., Savitski, M.M. & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem* (2012).

6. Hyotylainen, T. Novel methodologies in metabolic profiling with a focus on molecular diagnostic applications. *Expert Rev Mol Diagn* **12**, 527-538 (2012).

7. Ramisetty, S.R. & Washburn, M.P. Unraveling the dynamics of protein interactions with quantitative mass spectrometry. *Crit Rev Biochem Mol Biol* **46**, 216-228 (2011).

8. Schwamborn, K. Imaging mass spectrometry in biomarker discovery and validation. *J Proteomics* (2012).

9. Hui, P. Next Generation Sequencing: Chemistry, Technology and Applications. *Top Curr Chem* (2012).

10. Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486**, 215-221 (2012).

11. Yatsunenko, T., *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222-227 (2012).

12. Costello, E.K., Stagaman, K., Dethlefsen, L., Bohannan, B.J. & Relman, D.A. The application of ecological theory toward an understanding of the human microbiome. *Science* **336**, 1255-1262 (2012).

13. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214 (2012).

14. Dominguez-Bello, M.G., *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci U S A* **107**, 11971-11975 (2010).

15. Penders, J., *et al.* Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics* **118**, 511-521 (2006).

16. Gonzalez, A., *et al.* Our microbial selves: what ecology can teach us. *EMBO Rep* **12**, 775-784 (2011).

17. Stecher, B. & Hardt, W.D. Mechanisms controlling pathogen colonization of the gut. *Curr Opin Microbiol* **14**, 82-91 (2011).

18. Blumberg, R. & Powrie, F. Microbiota, disease, and back to health: a metastable journey. *Sci Transl Med* **4**, 137rv137 (2012).

19. Kosiewicz, M.M., Zirnheld, A.L. & Alard, P. Gut microbiota, immunity, and disease: a complex relationship. *Front Microbiol* **2**, 180 (2011).

20. Zhang, H., *et al.* Human gut microbiota in obesity and after gastric bypass. *Proc Natl Acad Sci U S A* **106**, 2365-2370 (2009).

21. Brown, C.T., *et al.* Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. *PLoS One* **6**, e25792 (2011).

22. Greenblum, S., Turnbaugh, P.J. & Borenstein, E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A* **109**, 594-599 (2012).

23. Saulnier, D.M., *et al.* Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. *Gastroenterology* **141**, 1782-1791 (2011).

24. Lemon, K.P., Armitage, G.C., Relman, D.A. & Fischbach, M.A. Microbiota-targeted therapies: an ecological perspective. *Sci Transl Med* **4**, 137rv135 (2012).

25. Brenchley, J.M., *et al.* High frequencies of polyfunctional HIV-specific T cells are associated with preservation of mucosal CD4 T cells in bronchoalveolar lavage. *Mucosal Immunol* **1**, 49-58 (2008).

26. Guadalupe, M., *et al.* Severe CD4+ T-cell depletion in gut lymphoid tissue during primary human immunodeficiency virus type 1 infection and substantial delay in restoration following highly active antiretroviral therapy. *J Virol* **77**, 11708-11717 (2003).

27. Brenchley, J.M., *et al.* CD4+ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. *J Exp Med* **200**, 749-759 (2004).

28. Lefrancois, L. & Puddington, L. Intestinal and pulmonary mucosal T cells: local heroes fight to maintain the status quo. *Annu Rev Immunol* **24**, 681-704 (2006).

29. Li, Q., *et al.* Peak SIV replication in resting memory CD4+ T cells depletes gut lamina propria CD4+ T cells. *Nature* **434**, 1148-1152 (2005).

30. Mattapallil, J.J., *et al.* Massive infection and loss of memory CD4+ T cells in multiple tissues during acute SIV infection. *Nature* **434**, 1093-1097 (2005).

31. Paiardini, M., Frank, I., Pandrea, I., Apetrei, C. & Silvestri, G. Mucosal immune dysfunction in AIDS pathogenesis. *AIDS Rev* **10**, 36-46 (2008).

32. Jiang, W., *et al.* Plasma levels of bacterial DNA correlate with immune activation and the magnitude of immune restoration in persons with antiretroviral-treated HIV infection. *J Infect Dis* **199**, 1177-1185 (2009).

33. Ancuta, P., *et al.* Microbial translocation is associated with increased monocyte activation and dementia in AIDS patients. *PLoS One* **3**, e2516 (2008).

34. Brenchley, J.M., *et al.* Microbial translocation is a cause of systemic immune activation in chronic HIV infection. *Nat Med* **12**, 1365-1371 (2006).

35. Mehandru, S., *et al.* Lack of mucosal immune reconstitution during prolonged treatment of acute and early HIV-1 infection. *PLoS Med* **3**, e484 (2006).

36. Guadalupe, M., *et al.* Viral suppression and immune restoration in the gastrointestinal mucosa of human immunodeficiency virus type 1-infected patients initiating therapy during primary or chronic infection. *J Virol* **80**, 8236-8247 (2006).

37. Cunningham-Rundles, S., *et al.* Effect of probiotic bacteria on microbial host defense, growth, and immune function in human immunodeficiency virus type-1 infection. *Nutrients* **3**, 1042-1070 (2011).

38. Anukam, K.C., Osazuwa, E.O., Osadolor, H.B., Bruce, A.W. & Reid, G. Yogurt containing probiotic Lactobacillus rhamnosus GR-1 and L. reuteri RC-14 helps resolve moderate diarrhea and increases CD4 count in HIV/AIDS patients. *J Clin Gastroenterol* **42**, 239-243 (2008).

39. Trois, L., Cardoso, E.M. & Miura, E. Use of probiotics in HIV-infected children: a randomized double-blind controlled study. *J Trop Pediatr* **54**, 19-24 (2008).

40. Gori, A., *et al.* Specific prebiotics modulate gut microbiota and immune activation in HAART-naive HIV-infected adults: results of the "COPA" pilot randomized trial. *Mucosal Immunol* **4**, 554-563 (2011).

41. Little, S.J., *et al.* Persistence of transmitted drug resistance among subjects with primary human immunodeficiency virus infection. *J Virol* **82**, 5510-5518 (2008).

42. Whitman, W.B., Coleman, D.C. & Wiebe, W.J. Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95**, 6578-6583 (1998).

43. Fox, S., Filichkin, S. & Mockler, T.C. Applications of ultra-high-throughput sequencing. *Methods Mol Biol* **553**, 79-108 (2009).

44. Chen, K. & Pachter, L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol* **1**, 106-112 (2005).

45. Baker, G.C., Smith, J.J. & Cowan, D.A. Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods* **55**, 541-555 (2003).

46. Chakravorty, S., Helb, D., Burday, M., Connell, N. & Alland, D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods* **69**, 330-339 (2007).

47. Youssef, N., *et al.* Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl Environ Microbiol* **75**, 5227-5236 (2009).

48. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**, R86 (2010).

49. Blankenberg, D., *et al.* Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* **Chapter 19**, Unit 19 10 11-21 (2010).

50. Giardine, B., *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451-1455 (2005).

51. Cole, J.R., *et al.* The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res* **35**, D169-172 (2007).

52. Sun, Y., *et al.* ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res* **37**, e76 (2009).

53. Wu, D., Hartman, A., Ward, N. & Eisen, J.A. An automated phylogenetic tree-based small subunit rRNA taxonomy and alignment pipeline (STAP). *PLoS One* **3**, e2566 (2008).

54. Dong, Q., *et al.* Diversity of bacteria at healthy human conjunctiva. *Invest Ophthalmol Vis Sci* **52**, 5408-5413 (2011).

55. Cunha, I.S., *et al.* Bacteria and Archaea community structure in the rumen microbiome of goats (Capra hircus) from the semiarid region of Brazil. *Anaerobe* **17**, 118-124 (2011).

56. Inceoglu, O., Al-Soud, W.A., Salles, J.F., Semenov, A.V. & van Elsas, J.D. Comparative analysis of bacterial communities in a potato field as determined by pyrosequencing. *PLoS One* **6**, e23321 (2011).

57. Ge, Y. & Wang, T.J. Circulating, imaging, and genetic biomarkers in cardiovascular risk prediction. *Trends Cardiovasc Med* **21**, 105-112 (2011).

58. Taly, V., Pekin, D., Abed, A.E. & Laurent-Puig, P. Detecting biomarkers with microdroplet technology. *Trends Mol Med* **18**, 505-516 (2012).

59. Noble, W.S. What is a support vector machine? *Nat Biotechnol* **24**, 1565-1567 (2006).

60. Kanwar, B., Favre, D. & McCune, J.M. Th17 and regulatory T cells: implications for AIDS pathogenesis. *Curr Opin HIV AIDS* **5**, 151-157 (2010).

61. Gupta, S.S., *et al.* Metagenome of the gut of a malnourished child. *Gut Pathog* **3**, 7 (2011).

62. Mendes, R., *et al.* Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332**, 1097-1100 (2011).

63. Kramski, M., *et al.* Novel sensitive real-time PCR for quantification of bacterial 16S rRNA genes in plasma of HIV-infected patients as a marker for microbial translocation. *J Clin Microbiol* **49**, 3691-3693 (2011).

64. Marchetti, G., *et al.* Microbial translocation predicts disease progression of HIV-infected antiretroviral-naive patients with high CD4+ cell count. *AIDS* **25**, 1385-1394 (2011).

65. Deeks, S.G., *et al.* Immune activation set point during early HIV infection predicts subsequent CD4+ T-cell changes independent of viral load. *Blood* **104**, 942-947 (2004).

66. Liu, Z., *et al.* Elevated CD38 antigen expression on CD8+ T cells is a stronger marker for the risk of chronic HIV disease progression to AIDS and death in the Multicenter AIDS Cohort Study than CD4+ cell count, soluble immune activation markers, or combinations of HLA-DR and CD38 expression. *J Acquir Immune Defic Syndr Hum Retrovirol* **16**, 83-92 (1997).

67. Shulzhenko, N., *et al.* Crosstalk between B lymphocytes, microbiota and the intestinal epithelium governs immunity versus metabolism in the gut. *Nat Med* **17**, 1585-1593 (2011).

68. Macal, M., *et al.* Effective CD4+ T-cell restoration in gut-associated lymphoid tissue of HIV-infected patients is associated with enhanced Th17 cells and polyfunctional HIV-specific T-cell responses. *Mucosal Immunol* **1**, 475-488 (2008).

69. Acheson, D.W. & Luccioli, S. Microbial-gut interactions in health and disease. Mucosal immune responses. *Best Pract Res Clin Gastroenterol* **18**, 387-404 (2004).

70. Bengmark, S. Use of some pre-, pro- and synbiotics in critically ill patients. *Best Pract Res Clin Gastroenterol* **17**, 833-848 (2003).

71. Sanders, M.E. Impact of probiotics on colonizing microbiota of the gut. *J Clin Gastroenterol* **45 Suppl**, S115-119 (2011).

72. Servin, A.L. Antagonistic activities of lactobacilli and bifidobacteria against microbial pathogens. *FEMS Microbiol Rev* **28**, 405-440 (2004).

73. Nissen, L., Chingwaru, W., Sgorbati, B., Biavati, B. & Cencic, A. Gut health promoting activity of new putative probiotic/protective Lactobacillus spp. strains: a functional study in the small intestinal cell model. *Int J Food Microbiol* **135**, 288-294 (2009).

74. Girardin, M. & Seidman, E.G. Indications for the use of probiotics in gastrointestinal diseases. *Dig Dis* **29**, 574-587 (2011).

75. Althaus, C.F.*, et al.* Rational design of HIV-1 fluorescent hydrolysis probes considering phylogenetic variation and probe performance. *J Virol Methods* **165**, 151-160 (2010).

76. Gianella, S.*, et al.* Associations between virologic and immunologic dynamics in blood and in the male genital tract. *J Virol* **86**, 1307-1315 (2012).

77. Hunt, P.W.*, et al.* T cell activation is associated with lower CD4+ T cell gains in human immunodeficiency virus-infected patients with sustained viral suppression during antiretroviral therapy. *J Infect Dis* **187**, 1534-1543 (2003).

78. Graham, N.M. The role of immunologic and viral markers in predicting clinical outcome in HIV infection. *AIDS* **10 Suppl 5**, S21-25 (1996).

79. Fahey, J.L.*, et al.* The prognostic value of cellular and serologic markers in infection with human immunodeficiency virus type 1. *N Engl J Med* **322**, 166-172 (1990).

80. Sonnenberg, G.F.*, et al.* Innate lymphoid cells promote anatomical containment of lymphoid-resident commensal bacteria. *Science* **336**, 1321-1325 (2012).

81. Vaishnava, S., *et al.* The antibacterial lectin RegIIIgamma promotes the spatial segregation of microbiota and host in the intestine. *Science* **334**, 255-258 (2011).

82. Hill, D.A. & Artis, D. Intestinal bacteria and the regulation of immune cell homeostasis. *Annu Rev Immunol* **28**, 623-667 (2010).

83. Slack, E., *et al.* Innate and adaptive immunity cooperate flexibly to maintain host-microbiota mutualism. *Science* **325**, 617-620 (2009).

84. Sandler, N.G., *et al.* Plasma levels of soluble CD14 independently predict mortality in HIV infection. *J Infect Dis* **203**, 780-790 (2011).