

ABSTRACT

Title of Dissertation:

IDENTIFICATION AND CHARACTERIZATION OF REGULATORY MIRNAS AND MRNAS IN THE LONGITUDINAL HUMAN HOST RESPONSE TO VAGINAL MICROBIOTA.

Steven Smith, Doctor of Philosophy, 2017

Dissertation directed by:

Dr. Jacques Ravel, School of Medicine,
University of Maryland, Baltimore.

The human vagina and the bacterial communities that reside therein exist in a finely balanced mutualistic association. Dysbiotic states of the vaginal microbiota, including bacterial vaginosis (BV), are characterized by a paucity of *Lactobacillus* spp., the presence of a wide array of strict and facultative anaerobes, and a pH >4.5. Symptoms such as odor and discharge can accompany these microbial dysbiotic states, however, epidemiologically, vaginal dysbioses have been associated with increased susceptibility to STIs, including chlamydia. The mechanisms by which vaginal microbiota protect or increase the risk to infections remain unknown. This thesis aimed to identify the molecular factors that control host cellular responses to *Lactobacillus* spp.-dominated and dysbiotic microbiota. Chapter 2 characterized the *in vivo* host microRNA (miRNA) response to different types of vaginal microbiota to gain insight into host functions that play a role in vaginal homeostasis. Leveraging daily collected vaginal samples in conjunction with a machine learning approach,

eight miRNAs were discovered to be differently controlled by vaginal microbiota. Of these, expression of miR-193b, known to regulate host cell proliferation, was increased by *Lactobacillus* spp.-dominated microbiota. *In vitro*, vaginal epithelial cells exposed to *Lactobacillus* spp. culture supernatants exhibited reduced epithelial cell proliferation, high miRNA-193b expression and decreased abundance of cyclin D1. More importantly, epithelial cell proliferation was identified as a requirement for efficient *Chlamydia trachomatis* infection. Chapter 3 characterized the *in vitro* transcriptome of epithelial cells exposed to *Lactobacillus* spp. relative to *Gardnerella vaginalis*, a surrogate for dysbiotic vaginal microbiota. Immune response and cell cycle pathways were found to be among the most modulated by *Lactobacillus* spp. Longitudinal gene expression suggested a role of histone deacetylases (HDAC) as an intermediary between immune stimulation and cell proliferation. Additionally, the epidermal growth factor receptor (EGFR), required for *C. trachomatis* infection, was decreased when epithelial cells were exposed to *Lactobacillus* spp. These findings contribute to the fundamental understanding of the vaginal microbiota's role in cellular homeostasis as a requirement for resistance to STI agents such as *C. trachomatis*, and ultimately will lead to improved preventive strategies against STIs through the modulation of vaginal microbiota composition and function.

IDENTIFICATION AND CHARACTERIZATION OF REGULATORY MIRNAS
AND MRNAS IN THE LONGITUDINAL HUMAN HOST RESPONSE TO
VAGINAL MICROBIOTA.

by

Steven Smith

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in
Biological Sciences
2017

Advisory Committee:

Professor Dr. Jacques Ravel, Chair
Professor Dr. Najib El-Sayed, Co-chair
Professor Dr. Michael Cummings
Professor Dr. Hector Corrada Bravo
Professor Dr. Mihai Pop

© Copyright by
Steven Smith
2017

Acknowledgements

I would like to first thank Jacques for his enthusiasm and continuous support. There was never an experiment or analysis that was too big, and I never felt judged or belittled for any ideas. I have become a better scientist, bioinformatician, thinker and writer because of his guidance and encouragement.

I thank my thesis committee members Drs. Najib El-Sayed, Hector Corrada Bravo, Michael Cummings and Mihai Pop for their valuable feedback, support and direction during this thesis. Their advice was internalized and my dissertation improved by it.

To all past and present Ravel lab members, I thank you for discussions on experimental design, hypothesis brainstorming, training, and support during my project. A special thank you to Beth Neuendorf and Vonetta Edwards who took time to train me on valuable skillsets in qPCR, aseptic technique and cell culture techniques. I thank Pawel Gajer for his statistical and computational advice. To Eli McComb, Lindsay Rutt, and Bilal Iqbal for keeping the lab lively and entertaining despite some stressful moments.

To my fellow graduate students and friends from the first year at College Park, I will always remember the memories we made teaching and experiencing graduate school. Our first year was key to maintaining our morale and I couldn't have gotten through it without you.

To my mom, Liz, dad, Scott and sister, Lauren, thank you for putting up with “endless school” that is now finally ending! You have been nothing but supportive

and curious. Even though dinner conversations were filled with words like ‘discharge’ and ‘odor’ and you never quite got the concept of microRNA or computational biology, I wouldn't have been where I am today if it wasn't for you.

To Kim, my fiancé, thank you for always listening to me ramble about science, my excitement and my frustrations. Thank you for making me care packages during quals, making dinners when I ran out of time and putting up with late night runs to the lab. When we first started dating I had no idea if my project was going to work. It was a big risk, but it paid off because now look, I am all done!

Table of Contents

Acknowledgements.....	ii
Table of Contents	iv
List of Tables.....	vii
List of Figures	viii
List of Abbreviations.....	x
Chapter 1 Introduction	1
The Role of the Vaginal Microbiota in Women’s Health.....	1
Characterization of the vaginal microbiota.....	1
The effect of the microbiota on host defense.....	7
Bacterial Vaginosis and Aerobic Vaginitis.....	10
Lactic acid and host defense.....	12
The effect of the microbiota on the vaginal mucosa.....	15
The effect of the microbiota on reproductive functions.....	16
Physiology affects vaginal microbiota composition.....	18
Concluding remarks.....	19
microRNAs As Regulatory Molecules to Host Responses.....	20
miRNA biogenesis and function.....	20
miRNA role in host response.....	25
Experimental Methods, Technology and Challenges of Measuring Human RNA.....	27
RNA handling and extraction.....	27
Bioinformatic and Computational Approaches to Identify Transcripts Associated with Experimental Conditions	30
Alignment of RNA-seq reads to reference and transcriptomic feature counting.....	30
Parametric methods to estimate statistically significant gene expression.....	30
Machine learning approaches to identify relevant genes.....	35
Concluding Remarks.....	39
Specific Aims.....	40
Specific Aim 1.....	40
Specific Aim 2.....	40
Chapter 2 microRNA expression induced by vaginal microbiota controls host epithelial cell proliferation and susceptibility to <i>Chlamydia trachomatis</i>	42
Introduction	42
Methods	44
Vaginal swab sample collection and metadata.....	44
Total RNA extraction.....	45
Small RNA sequencing library construction.....	47
Small RNA sequencing, quality control and read mapping.....	48
microRNA qPCR.....	49
Identification of BV-associated taxa and BV-associated miRNAs using Random Forest	50

Scratch assay using Bacterial Culture Supernatant (BCS).....	53
Cell Cyclin D1 (CCND1) Western blot.....	55
<i>C. trachomatis</i> infection and inhibition of cell proliferation	56
Results	57
Subject vaginal microbiota profile selection and small RNA-seq	57
Prediction of Amsel-BV diagnosis (proxy Amsel-BV) using metataxonomic data and metadata.....	60
Feature selection to identify miRNAs predictive of Nugent-BV and Proxy-Amsel-BV	63
Overexpression of miR-193b is associated with NBV	67
Epithelial cell proliferation decreased when exposed to <i>Lactobacillus</i> spp. Bacterial Culture Supernatants.....	72
Cell Cyclin D1 protein expression at 13 hours is reduced in vaginal epithelial cells exposed to <i>L. crispatus</i> and <i>L. jensenii</i> BCS but not <i>L. iners</i> BCS.....	75
<i>C. trachomatis</i> infectivity is reduced in non-proliferating cervical epithelial cells	79
Discussion	81
Chapter 3 <i>In vitro</i> vaginal epithelial cell transcriptional response to vaginal microbiota	89
Introduction	89
Methods	92
VK2 vaginal epithelial cell culture and bacterial culture supernatant (BCS) exposure .	92
Total RNA extraction from BCS exposed VK2 cells	93
Ribosomal RNA-depleted (rRNA-depleted) RNA sequencing library construction	94
RNA-seq library sequencing.....	96
Sequence trimming, alignment and feature counting.....	96
Read mapping quality control and differential expression	97
Pathway enrichment to identify commonly expressed pathways.....	98
Results	99
RNA-seq alignment statistics and quality control.....	99
Vaginal epithelial immune response and cell cycle pathways are associated with BCS exposure.....	100
Histone modification and cell cycle regulators are expressed dependent on BCS.....	106
Estrogen Receptor Alpha and Epidermal Growth Factor Receptor 1 gene expression are significantly decreased after <i>L. crispatus</i> and <i>L. jensenii</i> BCS exposure.....	110
Discussion	111
Concluding remarks.....	115
Chapter 4 Future work and conclusions	116
Future Work	116
Studying the microbiota-host dynamic in more complex <i>in vitro</i> cell systems	116
Further characterization of the role of miR-193b and mechanism of action.....	118
Evaluate the effect of HDAC inhibitors on cell proliferation, miR-193b expression and <i>C. trachomatis</i> infectivity	119
Elucidate the regulatory pathways leading to miR-193b expression	119
Conclusions	120
Appendices	125
Appendix 1 Steven Smith's contributions to thesis	125

Appendix 2 Scripts for Chapter 2 & 3 Analysis (R markdown file and rfSubjectSpecific.R)	129
Appendix 3 Cell Line Authentication Forms	272
<i>RESEARCH USE ONLY</i>	272
<i>RESEARCH USE ONLY</i>	273
<i>RESEARCH USE ONLY</i>	274
Appendix 4 Small RNA-seq miRNA raw read counts table.....	275
Appendix 5 Post-QC log2 normalized small RNA-seq counts and metadata.....	275
Appendix 6 Small RNA-seq library & subject metadata	275
Appendix 7 Metataxonomic data and metadata used to train and test the Amsel Random Forest model.....	275
Appendix 8 Importance metrics and p-values for the Amsel Random Forest variable selection results	276
Appendix 9 Metataxonomic data and metadata used as inputs to classify samples for proxy-Amsel diagnosis	276
Appendix 10 Importance metrics and p-values for the proxy-Amsel-RF and Nugent-RF variable selection results	276
Appendix 11 rRNA-reduced RNA-seq raw read counts table	276
Appendix 12 edgeR GLM-based LRT results.....	277
Bibliography	278

List of Tables

Table 2.1 Sample groups and number of samples pre- and post-sequencing QC, per subject	60
Table 2.2 Experimentally validated gene targets for each miRNAs found in proxy-Amsel-RF or Nugent-RF.....	65
Table 2.3 Gene Ontology processes for miR-193b based on the most number of experimentally validated targets.	67
Table 2.4 Tabular data for the qPCR time-course assay.....	69
Table 2.5 Tabular data for cell proliferation assay	75
Table 2.6 Tabular data for <i>C. trachomatis</i> infectivity assay.....	81
Table 3.1 Read alignment statistics for VK2 cells exposed to each BCS treatment for 4, 13, and 22h.....	101
Table 3.2 Number of differentially expressed genes per pairwise comparison.....	102
Table 3.3 Number of immune pathways (Figure 3.2) and cell cycle pathways (Figure 3.3) with absolute z-score greater than 2. Number of pro-inflammatory pathways within immune pathways in parenthesis.	103

List of Figures

Figure 1.1 Representative vaginal microbiota community state type generated by hierarchical clustering of microbial taxa composition and relative abundance.....	5
Figure 1.2 Representative longitudinal profiles of vaginal microbiota from women who self-sampled daily for 10 weeks.....	6
Figure 1.3 Typical BV and <i>Lactobacillus</i> spp. immune interactions in the vaginal epithelium.	11
Figure 1.4 Interaction of the human vaginal microbiota with physiology, host defense and reproduction.	13
Figure 1.5 Canonical microRNA biogenesis and mechanism.	21
Figure 1.6 RNA extraction, small-RNA-seq, ribo-reduced RNA-seq, and miR-qPCR.	28
Figure 2.1 Longitudinal profiles of vaginal microbiota for each subject in the study.	59
Figure 2.2 Quality Control (QC) figures of small RNA-seq samples.....	62
Figure 2.3 Random Forest variable importance ranking for proxy-Amsel-RF and Nugent-RF models and miRNA expression of the eight miRNAs identified by both models.	64
Figure 2.4 Gene Ontology processes of common significant miRNAs based on their experimentally validated gene targets.....	66
Figure 2.5 VK2 epithelial cell miR-193b expression time-course.	68
Figure 2.6 Relative expression of miR-193b following 4h exposure of D- or L-lactic acid.....	71
Figure 2.7 VK2 cell proliferation scratch assay, quantified by filled scratch area.	74
Figure 2.8 VK2 cell proliferation scratch assay, quantified by EdU (following page)	75
Figure 2.9 VK2 Cell proliferation scratch assay following 13-hour exposure to NYC-III (<i>Lactobacillus</i> spp.) or TSB (<i>G. vaginalis</i>) culture medium.	77
Figure 2.10 Western Blot of CCND1 after 13h exposure to BCS.....	78
Figure 2.11 Effects of cell proliferation inhibition on <i>C. trachomatis</i> infectivity in A2EN human cervical cells.	80
Figure 2.12 Vaginal epithelial cell homeostasis.	84
Figure 3.1 RINe distribution for RNA samples used in the study.....	99
Figure 3.2 Heat map of activation scores (z-score) from pathways associated with immunity for cells exposed to each BCS for 4h, 13h or 22h vs. cell culture medium.	104
Figure 3.3 Heat map of activation scores (z-score) from pathways associated with the cell cycle for cells exposed to each BCS for 4h, 13h or 22h vs. cell culture medium.	105
Figure 3.4 Longitudinal relative expression patterns for IL6 and IL8.....	106
Figure 3.5 IL6/IL1 signaling pathways gene expression after exposure to <i>L. crispatus</i> BCS for 13h compared to cell culture medium.	107
Figure 3.6 Cyclins and Cell Cycle Regulation pathway for averaged logFC of <i>L. crispatus</i> BCS and <i>L. jensenii</i> BCS 13h after exposure.....	108
Figure 3.7 Longitudinal relative expression patterns for selected cell cycle and chromatin remodeling genes.	109

Figure 4.1 Summary of Chapter 2 findings	121
Figure 4.2 Summary of Chapter 3 findings	123

List of Abbreviations

AKR1C2	aldo-keto reductase family 1 member C2
AMP	antimicrobial peptide
AV	aerobic vaginitis
BV	bacterial vaginosis
CART	classification and regression trees
CCL	chemokine (C-C motif) ligand
CCND1	cell cyclin D1
CCNE2	cell cyclin E2
CDC	cholesterol-dependent cytolysin
CDK4	cyclin dependent kinase 4
CDKN1	cyclin dependent kinase inhibitor 1A
CPM	counts per million
CST	community state type
EdU	5-ethynyl-2'-deoxyuridine
EGFR1	epidermal growth factor (EGF) receptor R1
ESR1	estrogen receptor- α
ETS1	ETS proto-oncogene 1 transcription factor
FC	Fold change
FDR	false discovery rate
GLM	generalized linear model
GO	gene ontology
HAT	histone acetyltransferase
HBD	human β -defensin
HDAC	histone deacetylase
HE4	human epididymis protein 4
HIV	human immunodeficiency virus
HNP	human neutrophil peptide
HPV	human papillomavirus
HSV	herpes simplex virus
IFN	interferon
IgA	immunoglobulin A
IgG	immunoglobulin G
IL	interleukin
IPA	ingenuity pathway analysis
KRAS	KRAS proto-oncogene, GTPase
logFC	log ₂ -transformed fold change
MAPK	mitogen-activated protein kinase
MAX	MYC associated factor X
MBL	mannose-binding lectin
miRNA	microRNA
MMP-8	matrix metalloproteinase 8
mRNA	messenger RNA
NBV	negative BV

NF-kB	nuclear factor kappa B
NF1	neurofibromin 1
NOD	nucleotide-binding oligomerization domain
Nugent-RF	Nugent score Random Forest model
OOB	out of bag
PBV	persistent BV
PE	paired end
PLAU	urokinase-type plasminogen activator
proxy-Amsel-RF	proxy Amsel diagnosis Random Forest model
PRR	pattern recognition receptor
Ribo-reduced	ribosomal RNA-reduced
RINe	RNA integrity number (e)
RNA	ribonucleic acid
RPKM	reads per kilobase per million mapped reads
rRNA	ribosomal RNA
SCFA	short chain fatty acid
SE	single end
SHMT2	serine hydroxyl transferase
sIgA	secretory immunoglobulin A
SLPI	secretory leukocyte peptidase inhibitor
SMCT1	solute carrier family 5 member 8
SP	surfactant protein
STI	sexually transmitted infection
TBV	transitional BV
TLR	toll-like receptor
TMM	trimmed mean of M-Values
TNF	tumor necrosis factor
UTR	untranslated region
YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta

Chapter 1 Introduction

*The Role of the Vaginal Microbiota in Women's Health*¹

Characterization of the vaginal microbiota

The assemblages of microbes (microbiota) associated with the human body have been shown to affect human physiology, immunity and nutrition [1]-[4]. In the vagina, microbes exist in a finely tuned mutualistic relationship with the host and provide the first line of defense against the colonization by opportunistic pathogens. Throughout a woman's lifespan, the vaginal microbiota undergoes major changes associated with transitional reproductive periods such as puberty and menopause [5]. During these periods, the vaginal microbiota can affect host reproductive physiology but can also be affected by host physiology.

Recent high-throughput 16S rRNA gene sequencing studies examining vaginal bacterial species composition and abundance in reproductive-aged women have shown that there are at least five major types of vaginal microbiota called community state types (CST) [6]-[8]. Four of these CSTs are dominated by either *Lactobacillus crispatus* (CST-I), *L. iners* (CST-III), *L. gasseri* (CST-II) or *L. jensenii* (CST-V) and one, CST-IV, does not contain a significant number of *Lactobacillus* spp. but is comprised of a polymicrobial mixture of strict and facultative anaerobes

¹ This section adapted from Smith, S. B. and Ravel, J. (2017), *The vaginal microbiota, host defence and reproductive physiology*. J Physiol, 595: 451–463. doi:10.1113/JP271694

including species of the genera *Gardnerella*, *Atopobium*, *Mobiluncus*, *Prevotella* and other taxa in the order *Clostridiales* (Figure 1.1) [6], [7], [9], [10]. The frequency of these CSTs has been shown to differ in different ethnic backgrounds, with CST-IV more common (~40%) in black and Hispanic women [6]. The polymicrobial condition known as bacterial vaginosis (BV) is compositionally similar to CST-IV since it is defined by a loss of *Lactobacillus* spp., the presence of anaerobes and strict anaerobes, and sometimes associated clinical symptoms including discharge, odor and irritation. In research settings, a Gram-staining scoring procedure that relies on the identification of bacterial morphotypes known as the Nugent test is used to establish a BV diagnosis [11] [appropriately renamed Nugent-BV by [12]]. Clinically, the diagnosis of BV is accompanied by an evaluation of the following signs and symptoms: discharge, malodor, the presence of clue cells and vaginal pH>4.5 as defined by the Amsel criteria [13].

Daily fluctuations in the composition of the vaginal microbiota have been previously documented by microscopy and cultivation studies [14]-[16]. These findings were confirmed and extended in longitudinal culture-independent analyses such as those of women who self-collected vaginal swabs twice weekly for 16 weeks [7], [17], [18], or daily for 10 weeks [19] or 4 weeks [20]. It was observed that some vaginal microbial communities transitioned in and out of CST-IV while others remained relatively stable either within a *Lactobacillus* spp.-dominated CST or CST-IV (Figure 1.2). The amount of time spent in a particular CST could vary individually as some women experienced consistent and stable CST longitudinal patterns (defined as community class), while others frequently transitioned between CSTs and most

frequently to CST-IV [7], [19]. In some cases, CST transitions were triggered by menstruation or sexual behaviors, but in other cases they seem to be driven by uncharacterized factors [7]. These longitudinal studies highlight the highly dynamic nature of vaginal microbial communities and emphasize the need to better understand the underlying biological factors modulating fluctuations in composition and functions that affect host physiology.

Historically, *Lactobacillus* spp.-dominated vaginal microbial communities have been associated with healthy reproductive-aged women and are characterized by the production of copious amounts of lactic acid and thus an acidic pH (<4.5) (reviewed in [21]-[23]). This acidic environment is thought to be highly protective against infections or colonization of the vagina by pathogens and non-indigenous microbes. An additional benefit of *Lactobacillus* spp. (i.e., *L. crispatus* and *L. gasseri*) is the supply of bacteriocins or bacteriocin genes (i.e., gassericin T, acidocin IF221A, type-A lantibiotic, and Bacteriocins IIa, IIc and J46) to inhibit growth of undesirable species (i.e., *Klebsiella* spp., *Staphylococcus aureus*, *Escherichia coli* or *Enterococcus faecalis*) [24], [25]. However, the notion that a *Lactobacillus* spp.-dominated vaginal microbiota is necessarily the norm has been called into question since mounting evidence suggests that about 25% of asymptomatic women do not possess a *Lactobacillus* spp.-dominated microbiota at any given time, a staggering proportion which does not support a diseased state [6], [26], [27]. These differences between women appear to be driven by a combination of cultural, behavioral, genetic, and other unknown underlying factors [6], [7], [26]. However, a strong association between CST-IV (as established by Nugent-BV) and increased risk to sexually

transmitted infections [28]-[30], including human immunodeficiency virus (HIV) [31]-[34], and reproductive tract and obstetric sequelae has been established through thorough epidemiological studies [35]-[39]. Hence, while CST-IV might be normal (asymptomatic) in some women, it is still associated with significantly increased risk to adverse outcomes. An illustration of how CST-IV can help further foster infections is in the case of chlamydial infection, where interferon- γ (IFN- γ) production is thought to be critical for chlamydia clearance. IFN- γ activates the human enzyme indoleamine 2,3-dioxygenase, which catabolizes tryptophan, eventually leading to tryptophan starvation and chlamydia clearance since genital chlamydia cannot synthesize tryptophan. However, production of indole compounds by anaerobes and strict anaerobes comprised in CST-IV affords chlamydia to shunt its deficiency and produce tryptophan, thus by-passing this host defense mechanism and establishing a long-term infection (discussed in [40]). Similarity, relative to other *Lactobacillus* spp.-dominated community states, CST-IV-like communities increase the risk to HIV infection [26], [41]. However, not all *Lactobacillus* spp. are necessarily beneficial and protective since, for example, some strains of *L. iners* might carry pathogenicity factors, such as inerolysin, a cholesterol-dependent cytolysin (CDC) and a host epithelial cell pore-forming enzyme, which was found to be up-regulated at least 6-fold in women with BV [42], [43]. Therefore, when considering the impact of the microbiota on host defense and reproductive physiology, it is important to place it in the context of these dynamic and individualized relationships.

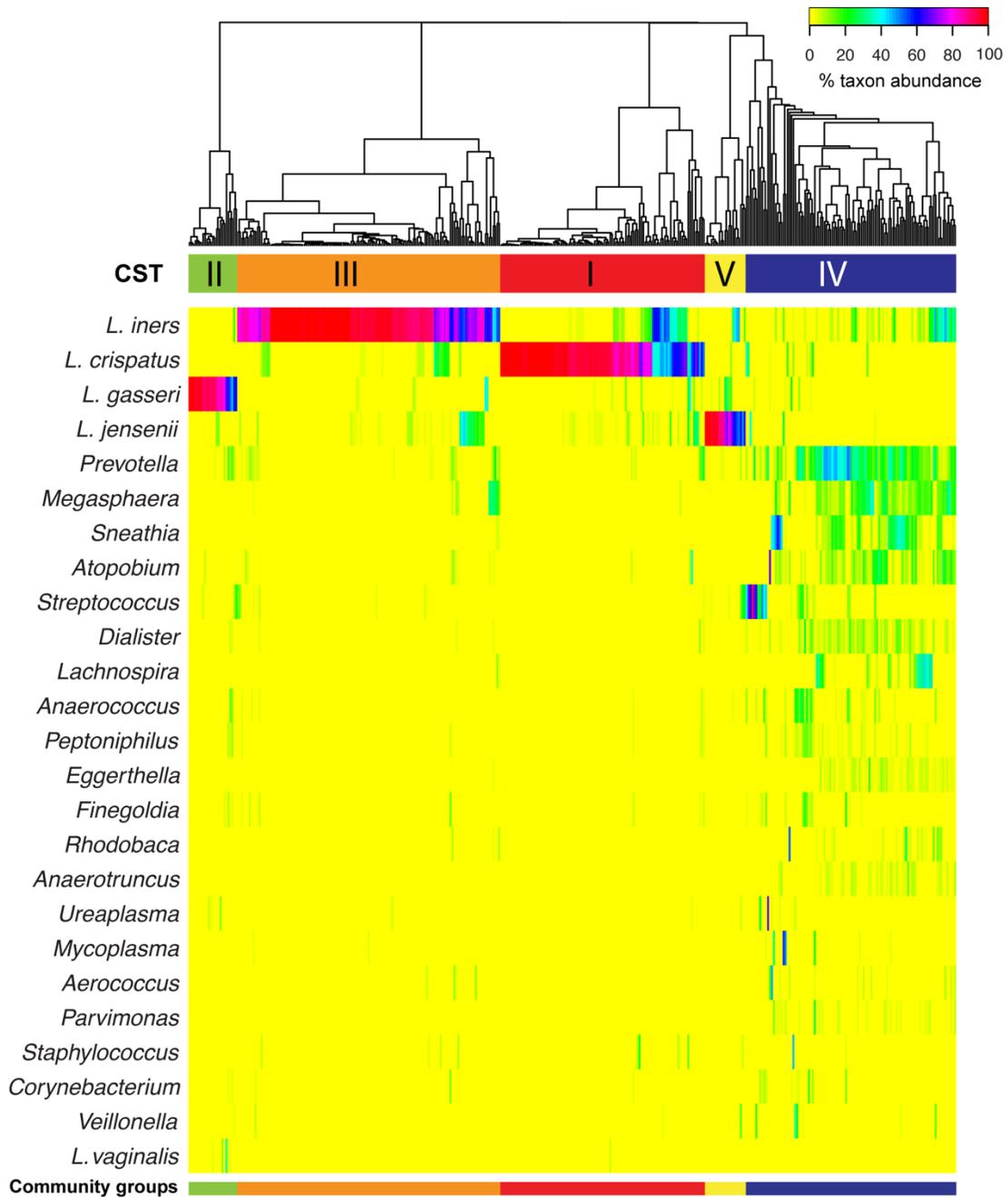


Figure 1.1 Representative vaginal microbiota community state type generated by hierarchical clustering of microbial taxa composition and relative abundance.

Hierarchical clustering shows that the vaginal microbiota of reproductive-aged women clusters into at least five distinct community state types, four of which are dominated by *Lactobacillus* spp. (*Lactobacillus crispatus* (CST-I), *L. iners* (CST-III), *L. gasseri* (CST-II) or *L. jensenii* (CST-V)) and the fifth (CST-IV) is comprised of a polymicrobial mixture of strict and facultative anaerobes including species of the genera *Atopobium*, *Megasphaera*, *Mobiluncus*, *Prevotella* and other taxa in the order *Clostridiales*. Figure reproduced from [6]

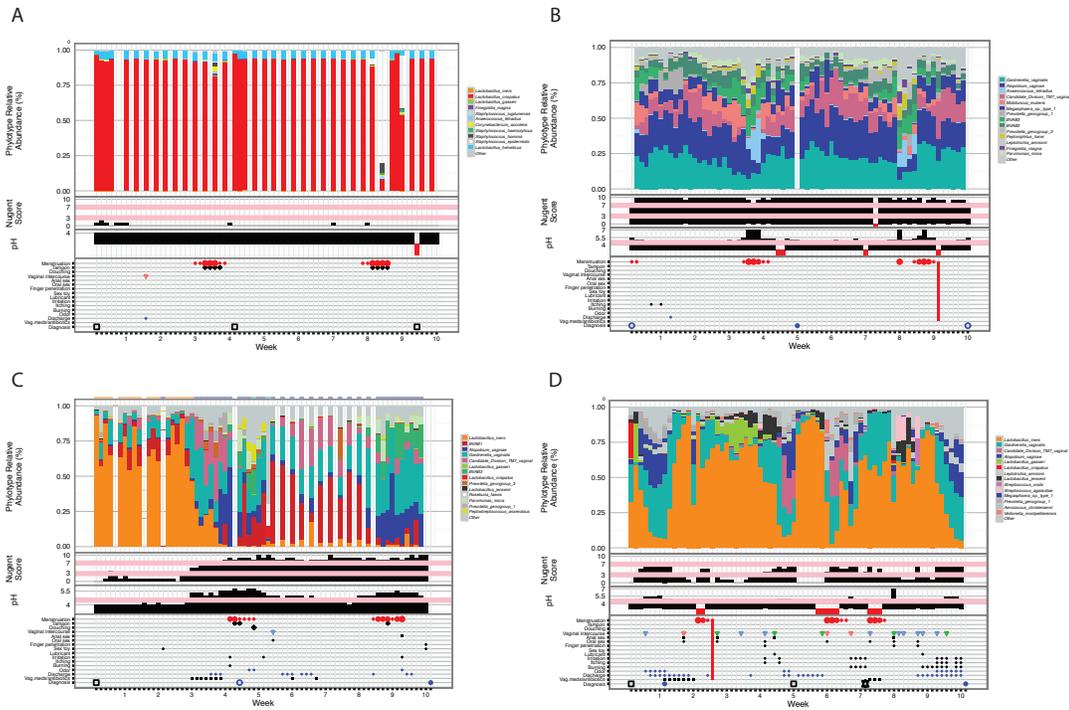


Figure 1.2 Representative longitudinal profiles of vaginal microbiota from women who self-sampled daily for 10 weeks.

Longitudinal plots of four participants over a 10-week study [19] who are either (A) consistently CST-I, (B) consistently CST-IV and (C-D) transition between CST-IV and CST-III. From top to bottom, panels show the profile of relative bacterial abundance based on 16S rRNA gene sequencing, Nugent score (0 to 10), pH (4 to 7) and metadata including reported menstruation (red dots), tampon use, douching, vaginal intercourse, anal and oral sex, finger penetration, sex toy use, lubricant use, vaginal irritation, vaginal itching, vaginal burning, vaginal odor, vaginal discharge, medication (e.g. antibiotics) and finally BV clinical diagnosis at approximately week 1, 5, and 10 (square is no clinical finding, open circle is asymptomatic Amsel-BV, filled circle is symptomatic Amsel-BV, star is UTI). Horizontal red lines in Nugent score plot indicate Nugent score 3 or 7 and lines in pH plot indicate pH 4.5. Red bars in Nugent score, pH and metadata plots denote missing data.

The effect of the microbiota on host defense

The vagina contains a number of immune-related cells and receptors to help sense the microbial environment [44]. Surveillance for microbes within the female genital tract of both commensal and pathogenic microbes is generally achieved by pattern recognition receptors (PRRs) such as Toll-like Receptors (TLRs) or the Dectin-1 receptor (which helps recognize the fungal pathogen *Candida albicans* [45], [46]), and nucleotide-binding oligomerization domain (NOD) receptors present in and on both squamous epithelial cells lining the vagina and the columnar cells lining the upper female genital tract (as reviewed in [44], [47]-[49]). Microbial stimulation of PRRs initiates cytokine/chemokine signaling cascades, for example secretion of IL-1 β , IL-6, IL-8 and tumor necrosis factor- α (TNF- α), to recruit/activate specialized cells, such as NK cells, macrophages, CD4+ helper T-cells, and CD8+ cytotoxic T-cell lymphocytes and B lymphocytes as reviewed in [44], [50], [51]. Genetic variants of PRRs such as the IL-1R antagonist, TLR4, TLR9, IL-1R2, or TNF- α may play a role in how a woman responds to a particular microbial challenge or pregnancy outcome, as evidenced by several genetic-disease association studies [52]-[58]. Women with CST-IV states show significant increases in IL-1 α , IL-1 β , TNF- α , IFN- γ , IL-4, IL-8, IL-10, IL-12p70, and fms-like tyrosine kinase 3 ligand relative to CST-I as well as significantly higher IFN- γ in CST-III relative to CST-I. Specifically, in one study, *Prevotella amnii*, *Mobiluncus mulieris*, *Sneathia amnii*, and *Sneathia sanguinegens* (all commonly found in CST-IV) were found to induce higher levels of IL-1 α , IL-1 β , and IL-8 relative to *L. crispatus* dominated communities (CST-I), whereas *L. iners* dominated communities (CST-III) induced moderate IL-8 levels

relative to CST-I. The authors also showed how there were significant increases in IL-1 α , IL-1 β and TNF- α longitudinally in subjects that transition from a CST-I, to CST-III and to a CST-IV [26]. Conversely, mock communities dominated by *L. crispatus* (CST-I) and *L. jensenii* (CST-V) on reconstructed three-dimensional vaginal epithelial models do not strongly elicit cytokine IL-1 β or IL-8 secretion relative to medium control, and also inhibit some pro-inflammatory responses after TLR 2/6 and 3 agonist induction [59]. These studies continue to support the notion that the innate immune response is largely driven by vaginal bacterial community states, with CST-IV potentially having a larger pro-inflammatory response than CST-I or CST-II, and with CST-III triggering an intermediate response.

Additional factors contributing to vaginal defense include mannose binding lectin (MBL), vaginal antimicrobial peptides (AMPs) and immunoglobulin A and G (IgA, IgG). As its name suggests, MBL binds mannose, N-acetylglucosamine and fucose carbohydrate moieties present on microbial cell surfaces. Eventually, this interaction leads to cell lysis or targeting for the immune system [60], [61]. IgA and IgG may help to prevent vaginal epithelial cell adherence and uptake, as well as contribute to the neutralization and clearance of infectious microbes from the vagina [62], [63]. Vaginal AMPs exist in various classes and may recruit immune cells via chemotaxis or possess anti-endotoxin activity. Mechanisms for each AMP have been thoroughly reviewed elsewhere [64]-[66], and while the specific association between AMPs and vaginal microbiota has not been extensively investigated, key findings are emphasized here. Defensins are a class of cationic and amphipathic AMPs with diverse mechanisms of action against common vaginal bacteria, pathogens and

viruses including HIV, herpes simplex virus (HSV) and human papillomavirus (HPV). In organotypic models of the vaginal epithelium, human α -defensin-2 (HBD-2) expression, but not that of HBD-1, was associated with colonization by *L. iners*, *Atopobium vaginae* and *Prevotella bivia* [67], while in another study using similar experimental *in vitro* conditions, *L. jensenii* but not *Gardnerella vaginalis* were shown to induce HBD-2 transcription [68]. As expected, many human defensins bind to viral-specific proteins to prevent viral attachment to a human cell surface, as for example, with retrocyclin-1, retrocyclin-2, human neutrophil peptide-1 (HNP-1), HNP-2, HNP-3 and to a much lesser degree HNP-4 [69]-[74]. In addition to defensins, other AMPs are found in the human vagina and include the secretory leukocyte protease inhibitor (SLPI), human epididymis protein 4 (HE4), LL-37, and Surfactant protein A (SP-A) and Surfactant protein D (SP-D). SLPI expression is associated with BV organisms [75] but not with *L. crispatus*, *L. iners*, *A. vaginae* or *P. bivia* [67], [76]. HE4 is associated with *G. vaginalis* [76] and LL-37 inactivates the sexually transmitted pathogen *Neisseria gonorrhoeae* while having no effect on *L. crispatus*, *L. jensenii* and comparatively little effect on *L. iners* [77]. The lack of AMP stimulation in response to some *Lactobacillus* spp. is associated with their needed maintenance in the vagina [78]. Similar to defensins, SP-A and SP-D contribute to viral inhibition, including HIV where they act via binding to the viral protein gp120 and human CD4, but with SP-A simultaneously enhancing gp120 binding to dendritic cells and therefore also facilitating HIV uptake [79], [80]. Thus, overall, microbes, environments, immune regulatory actions and genes tightly interact to govern homeostasis of the vaginal environment.

Bacterial Vaginosis and Aerobic Vaginitis.

As mentioned, a vaginal microbiota can be characterized by one of five CSTs, with CST-IV lacking a relatively high abundance of *Lactobacillus* spp. Generally, CST-IV can clinically manifest as aerobic vaginitis (AV) or BV, so the immune response to CST-IV outlined above overlaps considerably with BV or AV. AV is mainly differentiated from BV by the presence of an inflammatory response predominately associated with aerobes, such as group B *Streptococcus*, *Staphylococcus aureus*, *Escherichia coli*, and *Enterococcus* [81], [82]. The AV inflammatory response manifests symptomatically as itching or burning, molecularly as increased IL-6, IL-1 β , and cellularly as the presence of leukocytes or primary blood cells in a microscopic wet mount [83]. In contrast, the clinical definition of BV does not involve any overt inflammatory responses such as recruitment of neutrophils, redness, itching or burning (reviewed in [84]), but involves fishy odor and heavy discharge. A number of immune factors including IL-1 β , IL-2, IL-4, IL-6, IL-8, IL-10, IL-12, TNF- α , IFN- γ , chemokine C-C motif ligand 5 (CCL5) and SLPI have been variably and inconsistently associated with BV (summarized in [47]). These conflicting findings may be due to different study designs (longitudinal versus cross-sectional, or *in vitro* versus *in vivo*), different definitions of BV (symptomatic versus asymptomatic BV or Nugent-BV versus BV diagnosed according to the Amsel criteria) or that additional features actively suppress the inflammatory response in BV, such as IgA degradation, TLR expression inhibition, or immune-related genetic variants [85], [86]. As an example of BV's effect on host defense, cytokine analysis from a vaginal epithelial cell model co-colonized with mock communities

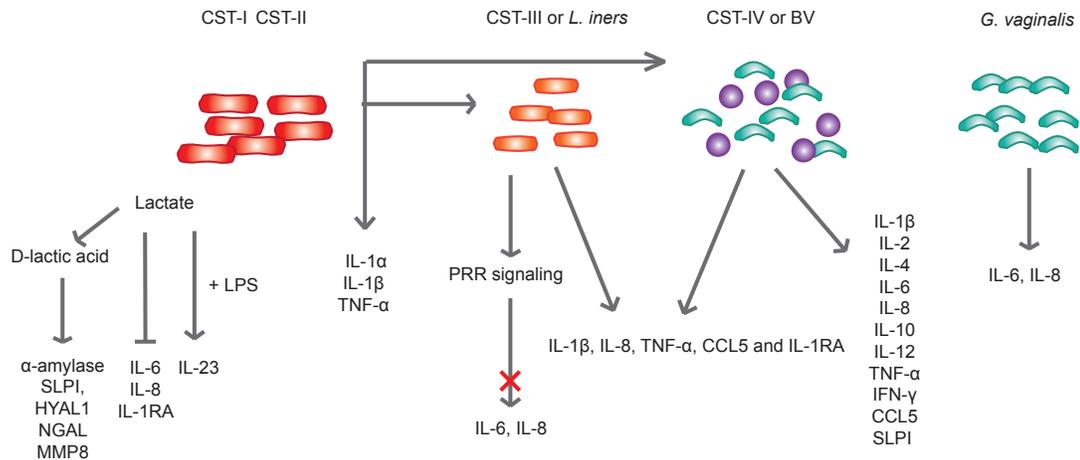


Figure 1.3 Typical BV and *Lactobacillus* spp. immune interactions in the vaginal epithelium.

Shown are immune responses associated with different vaginal microbiota CSTs. See text for details and abbreviations. Horizontal arrows from CST-I or CST-II to CST-III or CST-IV/BV indicate immune responses associated with transitions from/to these CSTs. Red shapes indicate either CST-I or CST-II microbiota, orange shapes indicate CST-III or *L. iners* while turquoise and purple represent *G. vaginalis* or CST-IV

representing CST-I to -IV as well as Nugent scores corresponding to respective BV diagnosis showed significant increases in IL-1β, IL-8, TNF-α, CCL5 and IL-1RA in CST-III or CST-IV, but not CST-I or CST-II (Figure 1.3) [41]. The ability of individual BV-associated bacterial species to elicit an *in vitro* immune response has also been studied, as in the cases of *A. vaginae* which induces expression of chemokine C-C motif ligand 20 (CCL20), HBD-2, IL-1β, IL-6, IL-8, and TNF-α via the NF-κB, TLR2 and MyD88 signaling pathways; *G. vaginalis* which induces IL-6, IL-8 transcripts; and *L. iners* which stimulates PRR signaling but not downstream inflammatory response cytokines IL-6, IL-8 or mucins [67], [87]. Bacteria-derived Short Chain Fatty Acids (SCFAs), namely acetate, butyrate, propionate and succinate, some of which exist at relatively higher proportions during BV, can induce a pro-inflammatory response under a hypothesis that SCFAs may act to ultimately inhibit chemotaxis and inflammation in BV [7], [88]-[91]. Relatively high concentrations (2-

20 mM) of acetate and butyrate, but not propionate, induce cytokine IL-6, IL-8 and IL-1 β secretions and also induce IL-8, TNF- α with TLR2 and TLR7 ligand stimulation in a dose- and time-dependent manner *in vitro* [90]. However, whether or not the host is actively downplaying the sensing of BV-associated microbes or a specific attribute of BV is evading inflammation remains to be demonstrated since the etiology of BV is still unknown and the necessary longitudinal studies are lacking [92].

Lactic acid and host defense

Lactic acid is produced mainly by vaginal microbes [93] and helps maintain healthy host physiological functions since it has been shown to directly inhibit *Chlamydia trachomatis* infection [94], and potentially both HSV-2 and HIV *in vitro* and *in vivo* if there is sufficient lactic acid to acidify the vagina to pH<4 (Figure 1.4) [95]-[98]. Lactic acid also inactivates a broad range of BV-associated microbes at pH<4.5 [99]. When *Lactobacillus* spp. dominate the vaginal microbiota, they acidify

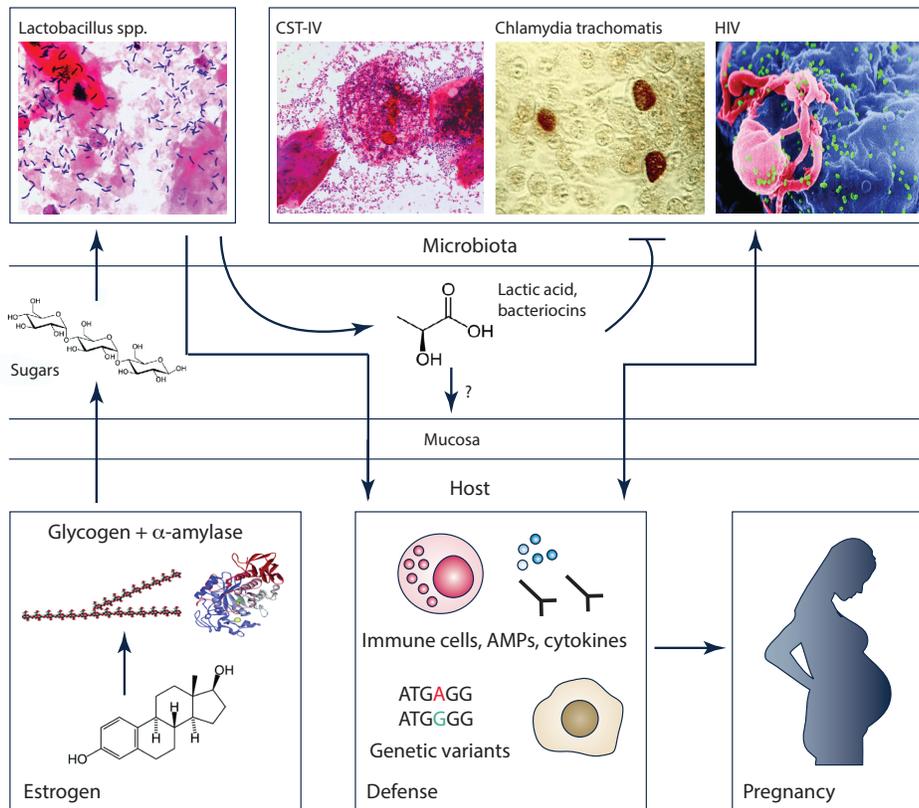


Figure 1.4 Interaction of the human vaginal microbiota with physiology, host defense and reproduction.

The host supplies glycogen and glycogen breakdown products via stimulation by estrogen production as a carbon source to the vaginal microbiota, therefore favoring *Lactobacillus* growth. In turn, lactic acid and bacteriocins produced mainly by *Lactobacillus* spp. contribute to the host defense while species associated with CST-IV, AV, or BV (such as *G. vaginalis* or *L. iners*) contribute to mucosa degradation and susceptibility to other infections. Specific host defenses including the innate immune response/inflammation, cell recruitment (such as NK cells or macrophages, pink and yellow circles, respectively), cytokines (pink “Y” shapes) and AMPs (blue circles) are dependent on host genetic polymorphisms, CST and pathogen presence (such as *C. trachomatis* and HIV) and may also have a negative effect on reproduction and pregnancy. Gram stain images of *Lactobacillus* spp. and CST-IV are unpublished data. All other images are in the public domain (Creative Commons License).

the vagina to a strongly acidic mean pH of 3.5 +/-0.2 that likely helps protect against a broad range of infections [89]. Recent studies aim to uncover the mechanism by which lactic acid can directly affect host immune functions, and showed that lactic acid can directly inhibit pro-inflammatory responses IL-6, IL-8 and IL-1RA, [100], induce the Th17 lymphocyte pathway via IL-23 in a dose-dependent manner upon lipopolysaccharide co-stimulation [101], and help release mediators from vaginal

epithelial cells and stimulate antiviral response by release of transforming growth factor- β [102]. In the gut, lactate and acetate from *L. casei* and *Bifidobacterium breve* inhibit cell proliferation, but whether these molecules play a similar role in vagina has not been studied [103]. Interestingly, lactic acid isomers may also play a role in determining host response and the subsequent host-microbiota relationship. Lactic acid exists in the vagina in both D(-) and L(+) isomers, with the host contributing only about 4-30% of the total lactate [93], suggesting a large reliance on microbes to supply the majority of lactic acid for protection. In one study, only D(-) lactic acid was correlated with α -amylase, SLPI, hyaluronidase-1 (HYAL1), neutrophil gelatinase-associated lipocalin (NGAL), and matrix metalloproteinase 8 (MMP-8) expression *in vitro* [75]. The authors suggest that epithelial cell exfoliation and subsequent breakdown of glycogen helps favor *Lactobacillus* spp. growth, and thus helps sustain D(-) lactic acid production (see discussion on α -amylase and glycogen below). Moreover, women with BV were found to be deficient in both isomers, while those with vulvovaginal candidiasis have elevated L(+) lactic acid as well as CD147 and MMP-8 genes [104]. *L. iners* does not produce D(-) lactic acid and fails to produce the L(+) lactic acid in abundance as high as *L. crispatus* or *L. gasseri* while *L. jensenii* produces only D(-) lactic acid [105], suggesting potential *Lactobacillus* species-specific effects on the host. Consequently, the composition of the vaginal microbiota, and specifically the ability of vaginal microbes to produce D(-) lactic acid, may help to inhibit pathogens and inflammatory responses while also favoring *Lactobacillus* spp. survival by using host cells resources for carbon sources.

The effect of the microbiota on the vaginal mucosa

The vaginal mucosa plays an important role as a physical barrier to separate host epithelial environment from harmful pathogens, including HIV [106], [107], whereas vaginal microbes can affect the integrity of the mucosa [108]. The BV-associated species *G. vaginalis* secretes sialidases, which have been shown to deglycosylate secretory immunoglobulin A (sIgA) and other sialoglycan substrates via the cleavage of sialic acid from glycoprotein at the α -2-3 and α -2-6 linkage Neu5Ac present on both *N*- and *O*-glycans, thereby hydrolyzing protective mucosal sialoglycoproteins [109]. *G. vaginalis* can consume and neutralize liberated sialic acid residues to further evade the host response [110]. In addition, *G. vaginalis* produces vaginolysin, a cholesterol-dependent cytolysin that could contribute to BV symptomology by forming pores in the vaginal epithelium with or without CD59 [111], [112]. BV-associated bacteria quantified by Nugent score also significantly associates with Mucins-1, -4, -5AC and -7 [113]. Conversely, certain types of vaginal communities could enhance the integrity of the mucosal barrier. A recent study showed that *L. crispatus*-dominated vaginal microbiota were able to reinforce the diffusional barrier properties of cervicovaginal mucus against HIV, hence hindering HIV penetration, while communities dominated by *L. iners* facilitated the penetration of HIV through the cervicovaginal mucus barrier [106]. Thus, a change in vaginal community composition and function is strongly associated with the integrity of the protective mucus layer, as such, vaginal bacteria, including species of *Lactobacillus*, can reduce or increase susceptibility to infectious agents, including HIV.

The effect of the microbiota on reproductive functions

The vaginal microbiota in combination with other factors is associated with adverse reproductive and obstetric outcomes. For example, a meta-analysis revealed that BV-like vaginal microbiota are significantly more prevalent in women with tubal infertility when compared to women with other causes of infertility, but is not associated with decreased conception rates [114]. Preterm labor and delivery has been thought to be in part associated with changes in the vaginal microbiota composition, namely bacteria found in CST-IV (i.e., *G. vaginalis* and *Ureaplasma*), AV or BV (where preterm labor and delivery are defined in each study as occurring before 37 weeks in either Caucasian women [115], mostly African American women [116], and mostly Caucasian women [117]). Jakovljević et al assessed differences of gestational time to delivery in BV versus non BV women, with a statistically significant difference of 37.72 ± 3.9 versus 39.59 ± 1.1 weeks [118]. However, another study did not find any association between CST-IV and preterm births (defined as 28–33.1 weeks versus term births of 38.8–40.7 weeks in mostly African American women) even though the study was well-powered and preterm births were phenotypically well-controlled [119]. It should be noted that differences in ethnicity, definition of preterm birth and analytical methods of microbiota data could explain these different observations. The possibility remains that functional differences exist and that hypotheses should be further explored using metagenomics or metatranscriptomics based approaches.

The interplay between host polymorphism and the vaginal microbiota could play an important role in the mechanisms by which microbes affect reproductive

health. Polymorphisms in genes that control inflammatory response (protein kinase C- α , fms-like tyrosine kinase 1, IL-6 and TNF- α) are associated with preterm delivery in combination with CST-IV vaginal microbiota, although the direct functional impact of these polymorphisms is unknown (Figure 1.4) [120], [121]. These studies and others have suggested that the inflammatory and antimicrobial peptide response, associated with certain vaginal microbiota, exhibit a role in rupturing and invading cervical plug or amniotic membranes, eventually triggering pro-inflammatory cascades that could lead to premature labor and delivery (reviewed in [64], [122], [123]). In a rhesus monkey model infected with group B *Streptococcus*, there was an observed increase in amniotic fluid of cytokines TNF- α , IL-1 β , and IL-6 occurring before uterine contractility or any clinical signs of infection, suggesting a direct role of infection in preterm labor [124]. Specific AMPs are expressed upon exposure *in vitro* and *in vivo* to the BV-associated bacteria *A. vaginae* (CCL20, HBD-2) or pathogens, such as *C. trachomatis* (elafin) and *N. gonorrhoeae* (SLPI), but not *L. crispatus* or *L. jensenii* [67], [87], [125]-[127]. The immune response to pathogens can trigger signaling cascades, which could ultimately lead to miscarriage, intrauterine infection, preterm labor, and tubal and ectopic pregnancy. For example, if *C. trachomatis* ascends past the cervix, such an infection in the upper genital track could lead to tubal scarring and potentially tubal infertility, ectopic pregnancy, and chronic pelvic pain ([128] and discussed in [129]). *C. albicans* triggers IL-6 secretion in the placenta ultimately leading to an increase in NF- κ B and an inflammatory response [130]. A study measuring the effects of immunomodulation in pregnant women with BV found that *C. albicans* and *Trichomonas vaginalis* but not *C.*

trachomatis or *N. gonorrhoeae* had no effect on vaginal cytokines, and none of these pathogens had any effect on anti-*G. vaginalis* hemolysin IgA, sialidase or prolidase activity [131]. Clearly there is still much to learn about the dynamics, the function and mechanisms driving the role of the vaginal microbiome in reproduction health.

Physiology affects vaginal microbiota composition

The composition of the vaginal microbiota changes throughout a woman's lifetime from birth, through puberty, reproductive age and menopause. At birth, *Lactobacillus* spp., if transferred from the mother, can colonize a baby girl vagina where the epithelium is mature and contain glycogen driven by the presence of circulating maternal estrogen [132]. After three to four weeks of age and until puberty, no estrogen is present or produced and the vaginal microbiota comprises a variety of anaerobes, diphtheroids, coagulase-negative staphylococci, and *E. coli*. Postmenopausal women often experience a loss of *Lactobacillus* spp. associated with the decrease in estrogen controlling vaginal epithelial proliferation, maturation, and accumulation of glycogen which is directly or indirectly nutritionally necessary for the maintenance of *Lactobacillus* spp. [133]-[137]. Indeed, estrogen levels peak during reproductive age and contribute to shaping the composition of the vaginal microbiota. In menopause, vaginal application of estrogen cream is associated with vaginal epithelial maturation, the accumulation of glycogen and acidic pH (<4.0), the latter indicative of the presence of high number of *Lactobacillus* spp. [138]. Interestingly, *Lactobacillus* spp. were originally thought to directly ferment glycogen in the vagina. However, this idea was gradually refuted and recent evidence suggests that human α -amylase catabolizes glycogen into smaller polymers, namely maltose

and maltotriose, which can then be used by *Lactobacillus* spp. for metabolism, even in newborns who have residual circulating maternal estrogen [139]-[143]. This model puts forward that the influence of estrogen, glycogen and especially α -amylase provides a positive selection pressure for a *Lactobacillus* spp.-dominated microbiota (Figure 1.4) [142]. These findings highlight the tight interplay between host physiology and the vaginal microbiota.

Concluding remarks

The human vaginal ecosystem is a dynamic environment in which microbes can affect host physiology but also where host physiology can affect the composition and function of the vaginal microbiota (Figure 1.4). Species of *Lactobacillus* have been historically associated with vaginal health in reproductive-age women due to the direct and indirect protective nature afforded by *Lactobacillus* products, such as lactic acid and bacteriocin among others, against mucus degradation and inhibition of pathogens. The reported inconsistent innate immune response observed with non-*Lactobacillus* spp.- or *L. iners*-dominated microbiota (CST-IV, BV, AV and CST-III, respectively), coupled with recent findings that question the definitions of normality, highlight the need for more in-depth functional understandings of the interaction between the vaginal microbiota and host physiology, reproduction and defense.

microRNAs As Regulatory Molecules to Host Responses

miRNA biogenesis and function

microRNAs (miRNA) are members of a class of non-coding regulatory RNA (ncRNAs) typically 20-22 nucleotides in length that have numerous roles in regulating biological processes via gene expression modulation, including organism development and innate immune signaling [144], [145]. Other ncRNAs include tRNA, rRNA, snoRNA, siRNA PIWI-interacting RNAs and long-non coding RNAs [146]. miRNA biogenesis, regulation and mechanisms have been studied using various model organisms including poriferans, cnidarians, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Gallus gallus*, *Rattus norvegicus*, *Mus musculus*, and *Homo sapiens* [144], [147], [148]. miRNA mechanisms in plants are largely distinct from metazoans and therefore not discussed [144]. Furthermore, where appropriate, human-specific mechanisms are only discussed.

The canonical metazoan miRNA biogenesis pathway begins in the cell's nucleus by RNA polymerase II-dependent transcription of pri-miRNA (Figure 1.5) [149], [150]. Pri-miRNAs possess both a 5' 7-methyl guanylate (m7G) cap and a 3' poly(A) tail [150] identical to those found in coding messenger RNA (mRNA). Pri-miRNAs exist as part of both long non-coding RNAs and in miRNA family clusters [151]. A bioinformatic analysis suggested some miRNAs are co-transcribed with their host genes while others are transcribed by themselves [152]. A non-exhaustive list of transcription factors that regulate miRNAs include tumor protein p53 (TP53),

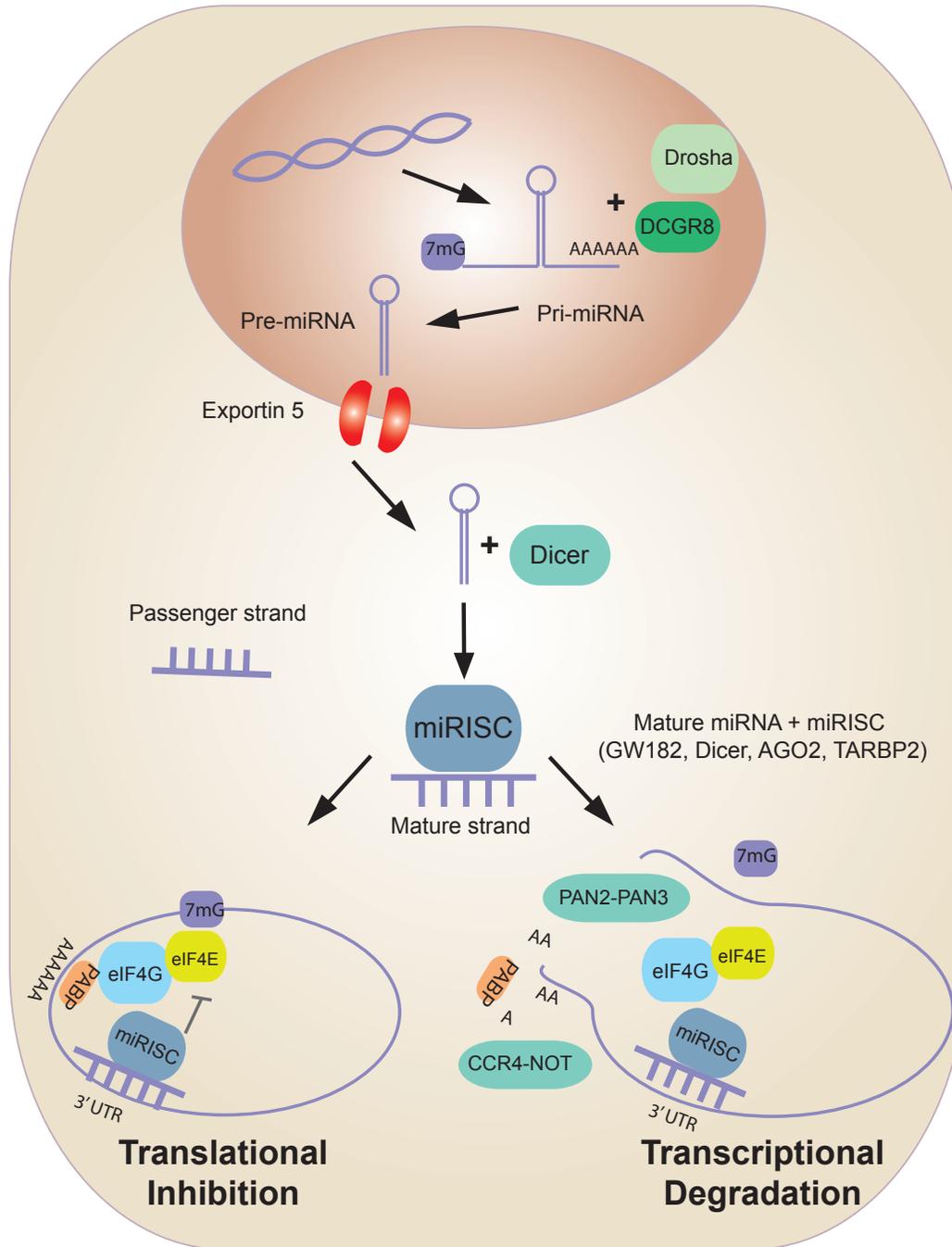


Figure 1.5 Canonical microRNA biogenesis and mechanism.

Pri-miRNAs are transcribed from DNA as transcripts containing a 7-methylguanosine cap and poly(A) tail. The hairpin structures that form on pri-miRNA are recognized by Drosha and DCGR8 which cleave the hairpin structure from the transcript, resulting in pre-miRNA. Pre-miRNA is exported from the nucleus by Exportin 5, then the loop structure is cleaved by Dicer. The mature strand is loaded onto miRISC which contains at minimum AGO2, GW182, Dicer and TARBP2. miRISC typically binds to the 3' UTR of a target mRNA and has been shown to repress protein translation by disrupting PABP from binding to eIF4G and other scaffold proteins including eIF4E. Transcriptional degradation has also been observed via mRNA de-capping and de-adenylase factors including CCR4-NOT and PAN2-PAN3, resulting in further destabilization of the transcript.

myelocytomatosis oncogene (MYC), myoblast determination protein 1 (MYOD1), zinc finger E-box binding homeobox 1 (ZEB1), and ZEB2, with additional pri-miRNA transcriptional control mechanisms still being an active research area [146]. Pri-miRNAs form hairpin structures after they are transcribed [11] and are recognized and cleaved by the nuclear proteins DCGR8 and the RNase-III like enzyme Drosha, initiating pre-miRNA processing and forming ~65 nucleotide hairpin structures [153]. Pre-miRNAs are exported from the nucleus into the cytoplasm by Exportin 5 requiring RanGTP [154]-[156].

Nucleus-exported pre-miRNAs are further processed into mature miRNA in the cytoplasm by a second RNase-III enzyme Dicer, which recognizes and cleaves a 22 nucleotide long dsRNA from both the 3' and 5' ends (Figure 1.5) [157], [158]. One of the two mature complementary hairpin strands (the guide strand) is loaded into the miRNA-Induced Silencing Complex (miRISC) comprised of the Argonaute 2 (AGO2) protein, GW182 (a 182 kD protein containing glycine (G) and tryptophan (W) repeats [159]), RNase-III like enzyme Dicer and human immunodeficiency virus trans activating response RNA-binding protein 2 (TARBP2) while the second strand (the passenger strand) is degraded [160] [161]. Characteristics that determine which strand becomes the guide strand include thermodynamic stability, 5' nucleotide identity and structure, and the presence of co-factors TARBP2 or protein activator of interferon induced protein kinase EIF2AK2 (PRKRA) [162]. Argonaute is not only critical to the miRNA biogenesis pathways, but was found to be fatal after deletion in mouse embryonic cells [163].

Eukaryotic protein translation of an mRNA proceeds in three stages: initiation, elongation and termination where each stage is highly orchestrated, consisting of many co-factors, ribosomal RNA and transfer RNA [164]. During initiation, the mRNA is circularized through the binding of the mRNA's 7mG cap to its poly(A) tail through binding and scaffold proteins including poly(A) binding protein (PABP) and eukaryotic translation initiation factor 4 gamma 1 (eIF4G) and eIF4E [165], [166]. Although an active area of research is the exact mechanisms, co-factors and kinetics by which translational inhibition and transcriptional degradation is achieved, there are two currently accepted mechanisms by which miRNAs exert their repression of target genes: translational initiation inhibition and mRNA degradation (Figure 1.5) [167]-[170]. Note these mechanisms are not mutually exclusive, and indeed, they are thought to occur almost simultaneously [171]. It has been demonstrated that translational inhibition precedes degradation, although interestingly in one study, mRNA destabilization explains 66% to ~90% observed repression and therefore may ultimately drive gene regulation [172], [173].

During translational initiation inhibition, the miRNA that is loaded into the miRISC binds to the semi-complementary sequence of an mRNA target molecule typically found in the 3' Untranslated Region (UTR) and is thought to disrupt translational machinery through the dissociation of the translational co-factors eukaryotic translation initiation factor 4AI (eIF4AI) and eIF4AII from the translation complex [174] and inhibiting circulation by repressing PABP and eIF4G binding [175], [176]. Interestingly, some miRNAs have been found to bind to the 5' UTR of certain genes and in some instances caused increased translation [38]-[42].

Additionally, AGO2 possesses motifs that resemble the binding domain of eIF4E to the m7G cap-binding domain suggesting that Argonaute is able to bind to the cap structure instead of eIF4E during translational initiation [177]. Depleting eIF6 in human cells, a factor involved in assembling the 80S ribosome, removed miRNA-mediated effects [178] suggesting it is required for effective inhibition.

miRNA-induced mRNA degradation has also been characterized in certain contexts. Components of the miRISC have been observed to recruit or associate with de-adenylase complexes and de-capping enzymes that destabilize or actively degrade mRNA, including the de-adenylase associated complexes CCR4-NOT and PAN2-PAN3 [179]-[182]. Studies suggest that the GW182-mediated recruitment of the CCR4-NOT complex helps release PABP from the poly(A) tail [183]. In *Drosophila melanogaster*, miRISCs promote de-capping through recruitment of Decapping protein 1 (DCP1), maternal expression at 31B (Me31B) and Protein associated with topo II related-1 (Patr-1) independently of de-adenylation [184]. One study found that Argonaute, GW182, CCR4-NOT and DEAD-box helicase 6 (DDX6)/Me31B repress and degrade polyadenylated mRNA targets during ribosome scanning independent of eIF4A [185].

Non-canonical miRNA biogenesis pathways have also been discovered. Some miRNAs originating from introns (mitrons) and the spliceosome machinery can be processed without Drosha, which then enter the canonical pathway beginning with Exportin 5 export [186], [187]. In another non-canonical biogenesis pathway, most of the miR-105 and let-7 families belong to miRNAs that have a 3' mono-uridylation added by TUTase, which is required for recognition by Dicer [188]. Drosha- and

Dicer-independent pathways and miRNAs have also been described but are believed to occur less frequently [189], while recent evidence has challenged the requirement of Dicer and Exportin 5 as miRNAs were found to be processed or loaded onto AGO2 but with reduced efficiency [189]. Indeed, vertebrate pre-miR-451 is Dicer independent and is instead cleaved by AGO2 [190]. Finally, miRNAs can exhibit either SNP or end modifications which can alter the stability of loading on RISC and of mRNA binding [191], [192].

miRNA role in host response

The host first recognizes microbes through the innate immune system [48], [193], [194]. As such, microbes trigger innate immune response pathways which lead to expression and regulation by immune-associated miRNAs. These miRNAs include miR-155, miR-146, miR-125, miR-223, miR-21 and the let-7 family which regulate immune functions such as the NF- κ b response, TLR receptor signaling, and IL-6, IL-1 β and IL-8 cytokine transcription. Interestingly, these immune responses are also potentially implicated in BV or vaginal dysbiosis [145], [195]-[200]. As the host must balance and regulate the immune response to pathogenic and mutualistic microbe, there is growing appreciation for the role that miRNAs might play in the host response to different microbiota in general.

The role of microbiota miRNA-mediated gene regulation has been studied in the mouse gut. In one study, differential expression of nine miRNAs in the ileum or the colon of germ free mice relative to pathogenic-free microbiota mice was found. In this system, the mouse gene *Abcc3* (a transporter protein) is regulated by *mmu-miR-665* [201]. Interestingly, the gut microbiota of Dicer-mutant mice, which broadly

impacts miRNA processing, show exacerbated colitis, which was eliminated by the addition of wild type fecal miRNA administered by gavage at 22.5 $\mu\text{g}/\text{day}$ for 7 consecutive days [202]. In Dicer-null gut epithelial cells, IL-13 induced miR-375, which targets Kruppel-like factor 5, believed to be a myeloid transcription factor, to maintain mucosal immunity [203], [204]. These studies point to critical roles for miRNAs in regulating the innate immune response, and the maintenance of both epithelial homeostasis and barrier function.

Certain bacteria in the mouse and human gut microbiota produce a relatively high abundance of the short chain fatty acid butyrate which has been shown to decrease cell proliferation via histone deacetylase (HDAC) inhibition and has also been shown to alter the expression of many RNA molecules within cell cycle pathways [205]-[209]. For example, butyrate decreased human colorectal carcinoma HCT-116 cell proliferation by stimulating the expression of cyclin dependent kinase inhibitor 1A (CDKN1A, also known as p21) in a dose dependent manner. This effect was dampened by adding exogenous miR-106b, which targets CDKN1A and thus reversed the inhibition of cell proliferation [210]. Although lactate is known to inhibit pro-inflammatory responses in vaginal cells and inhibit cell proliferation in gut cells, it is not known whether the same is true for the vaginal environment [100]-[102], [211]. Butyrate is transported into the cell via transporter solute carrier family 5 member 8 (SLC5A8) [212], [213]. More generally, SCFA have been also shown to inhibit inflammatory responses in various cell lines [214]. For example, butyrate inhibits innate immune responses through NF- κ B inhibition [215], [216]. Interestingly, acetic and butyric acid at relatively high concentrations (20 mM), in

conjunction with TLR ligands, were also found to induce pro-inflammatory cytokine responses in part through the release of reactive oxygen species in primary blood mononuclear cells (PBMCs) or neutrophils [90]. Thus, SCFAs released by bacteria in the microbiota play a critical role in host cell response, potentially through immune-related pathways and mediated through miRNA.

Experimental Methods, Technology and Challenges of Measuring Human RNA

RNA handling and extraction

Unlike DNA, RNA is highly susceptible to RNases which catalyze RNA degradation [217]. High quality RNA, evaluated in biological samples by the abundance of intact ribosomal RNA with the RNA Integrity Number (RINe) [218], is necessary for RNA sequencing as longer reads generate higher confidence sequence alignments to reference genomes. To achieve this goal, tissue and cell samples are stored in specialized transport media such as RNAlater [219] designed to inhibit RNases and stabilize RNA, prior to RNA extraction. Dedicated RNase-free laboratory bench space, equipment and precautions aimed at reducing RNase contamination from laboratory personnel's skin and clothing, is essential for obtaining consistent, high quality RNA [217]. RNA handling and extraction therefore involves additional labor-intensive processing and handling relative to that of DNA.

Once extracted and purified from a biological sample, RNA is typically characterized and quantified using high throughput sequencing or qPCR (Figure 1.6).

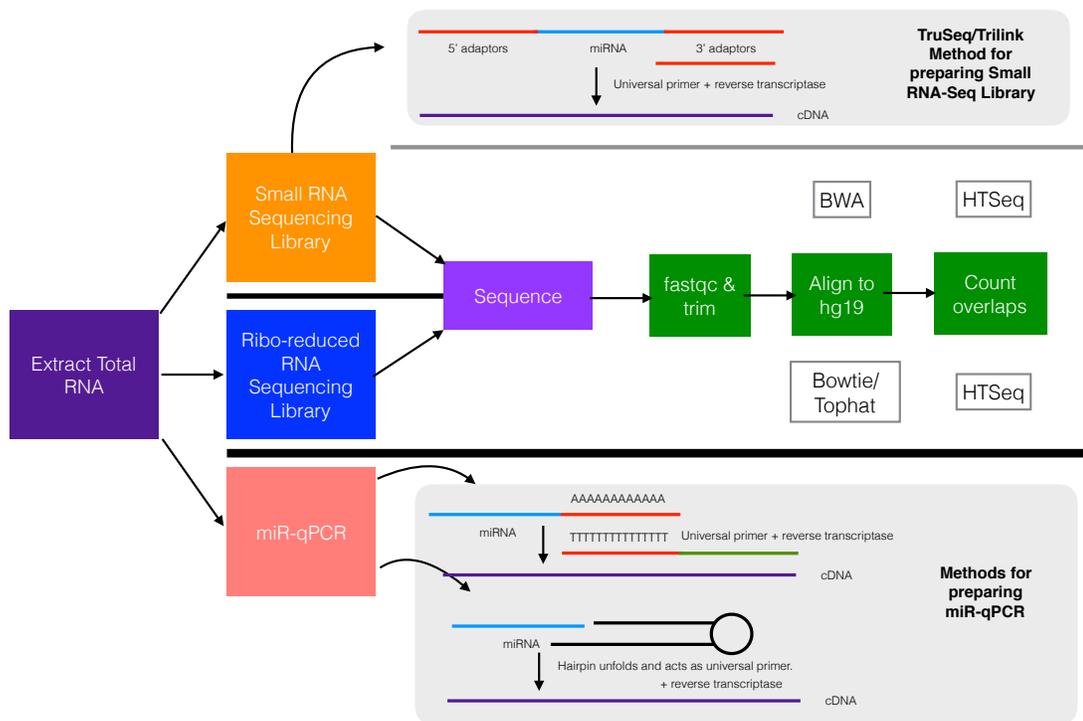


Figure 1.6 RNA extraction, small-RNA-seq, ribo-reduced RNA-seq, and miR-qPCR.

Total RNA is extracted from samples stored in RNAlater, then prepared for either small RNA-seq, ribo-reduced RNA-seq or miR-qPCR. Once sequencing libraries are constructed and sequenced, reads are aligned to a human genome sequence reference using BWA or Bowtie/TopHat, then genomic features are counted using HTSeq. Small RNA-seq libraries must first be ligated with sequencing adaptors before creating cDNA, which has the potential to produce adaptor dimers unless special adaptors are used (as is available commercially from TriLink). Similarly, miR-qPCR lengthens the target molecule by addition of a poly(A) tail or a hairpin probe so that reverse transcriptase can prime the miRNA.

Both approaches convert RNA into cDNA using the enzyme reverse transcriptase. For miRNA characterization, small RNA-seq or miRNA-qPCR is performed whereas with longer RNA transcripts (e.g. mRNA), RNA-seq or qPCR is performed. Due to their small size, reverse transcriptase cannot prime a miRNA template in the same manner as longer-length RNAs. Therefore, the template length must be increased before RNA is converted into cDNA. The first step in small RNA-seq library construction is ligation of 3' and 5' adaptor sequences that are used to prime a

universal sequence for the RNA reverse transcriptase to form cDNA. Thus, a major challenge of small RNA-seq is to ligate adaptors to small RNA targets and not to each other (which form adaptor dimers). Removal of adaptor-dimers by gel purification is often difficult as the dimers are close to the size of the library (only differing by 20 base pairs). A novel commercial method alleviates this problem by using proprietary chemically modified adaptors containing end-modified oligonucleotides that selectively bind to small RNA but not each other [220]. For qPCR, the two main approaches to increase effective template length are adding a poly(A) tail to miRNA to bind a universal poly(T) primer [221] [222] or using hairpin loop hybridization [223] [224]. Both approaches result in cDNA for qPCR.

Ribosomal RNA-reduced (ribo-reduced) RNA libraries are constructed by first removing abundant human rRNA molecules from the sample using beads that hybridize human rRNA [225]. The remaining RNA is reverse transcribed into cDNA using random hexamer or poly(T) priming. Sequencing adaptors are added in addition to multiplex indices and then sequenced [226]. Sequences produced through the process are aligned to the human reference to characterize RNA expression level as a function of gene coverage and depth of coverage.

Bioinformatic and Computational Approaches to Identify Transcripts Associated with Experimental Conditions

Alignment of RNA-seq reads to reference and transcriptomic feature counting

For both human small RNA-seq and human ribo-reduced RNA-seq, sequence reads are mapped to a human reference genome sequence using read mapping algorithms, such as BWA [227] or Bowtie/Tophat (Figure 1.6) [228]. Whereas BWA is a relatively fast short-read aligner designed to rapidly align sequence reads [227], aligners such as Tophat are gap-aware and incorporate splice-junctions [228]. BWA and TMM normalization (discussed below) were found to perform optimally with small RNA-seq read alignment when compared to Bowtie and other normalization methods [229]. Once sequence reads are aligned to the human reference genome sequence, the number of reads overlapping with genomic features (e.g. genes) are counted based on a known annotation and genomic coordinates, using, for example, the popular annotation tool HTSeq. [230]

Parametric methods to estimate statistically significant gene expression

The human transcriptome can be characterized and measured using RNA-seq to understand the transcriptional host response in a given clinical or experimental condition. A transcriptomic experiment aims to identify differentially expressed genes (defined as a genomic feature under study, e.g., an exon, isoform, or gene transcript) between sample groups. Differential expression is typically defined using both a relative gene expression difference between one or more conditions, (log-transformed fold change of normalized expression values) and a statistical test for significance of

the observed expression differences, where one or both values exceed a pre-determined threshold value. Relative expression values are usually calculated using a gene's mean estimated expression value in the experimental condition relative to a control, or using a Generalized Linear Model (GLM)'s coefficient term (detailed below). Statistical significance is typically calculated with a parametric test using each gene's expected value and variance estimates. In comparison to microarray data, in which observed gene expression takes on a continuous probe intensity value, RNA-seq data is expressed as discrete sequence read counts and is therefore modeled using a discrete random variable. Empirical data have demonstrated that counts from RNA-seq experiments fit a negative binomial distribution, which is a Poisson distribution (encapsulating technical variance) with additional variance due to biological replication [231]-[237]. Popular differential expression analysis tools that employ the Negative binomial distribution include DESeq [238], DESeq2 [239] and edgeR [233], [240] and model an observed RNA-seq count Y , from gene g and sample (library) i as:

$$Y_{gi} \sim NB(\mu_{gi}, \phi_{gi}) \quad (1)$$

$$\phi_{gi} = \mu_{gi} + \alpha_i \cdot \mu_{gi}^2$$

$$\mu_{gi} = \pi_i \cdot \lambda_{gi} \quad (2)$$

Where μ is the mean expression, ϕ is the variance, α is dispersion, λ is the number of sequence reads aligning to g and π is the scaling normalization factor.

Additionally, edgeR and DESeq2 use the log-link linear function of the scaled mean:

$$\log \mu_{gi} = \beta_{g0} + x_i \cdot \beta_{gT} \quad (3)$$
$$\begin{cases} x_i = 0 & \text{if } i \text{ in control group} \\ x_i = 1 & \text{if } i \text{ in treatment group} \end{cases}$$

The λ term in equation (2) represents the per-library and method-specific normalization scaling factor to correct observed sequence read counts for inter-sample differences in library size (sequencing depth), gene length and GC content, although most RNA-seq normalization methods only account for sequencing depth as the intra-gene length and GC content is assumed to remain constant between samples. Four relatively straightforward normalization methods include: 1) scaling reads by the number of reads per kilobase per million mapped reads for each sample (RPKM) [19]; or scaling all gene counts by the 2) total number of reads for each sample and multiplied by the mean total read count of all samples (total count, TC), or 3) non-zero upper quartile (UQ) of counts, or 4) median (Med) non-zero counts [241]. For example, the normalization method implemented in DESeq applies a correction factor to each gene within a sample based on the median ratio of each gene's read count relative to the geometric mean for all samples. The Trimmed Mean of M-Values (TMM) method, as implemented in edgeR, first trims the extreme read counts from the calculation and then computes read count log ratios for each gene against a reference sample (M-values) [242]. The weighted mean of M-values is applied as a correction factor to each sample. Accurate normalization is paramount to downstream analysis such that read count values capture biological instead of technical signal. In evaluation tests, RPKM and TC were found to be ineffective as gene length and differential gene expression varies and it does not remove bias [241], [243] [241].

Conversely, the UQ, Med, DESeq, and TMM methods performed similarly when assessed on results of differential expression analysis, with TMM and DESeq methods outperforming others with regard to maintaining a low false positive rate without loss of power [243].

The variance of the negative binomial distribution in equation (1), ϕ_i , is a function of the mean, μ_i and a scalar α known as the dispersion parameter which governs extra variation (biological variation). For $\alpha \ll 0$, the variance approaches μ_i and resembles the Poisson distribution [237]. A common task for differential expression algorithms is to estimate both the scaling parameter λ and dispersion parameter α for subsequent statistical inference testing. However, there are two related challenges when estimating dispersion: low number of within-group replicates and heteroskedasticity of RNA-seq counts. The first challenge is that of skewed dimensionality, termed ‘the curse of dimensionality’, wherein a typical experiment has a high number of genes relative to the number of replicates within a sample group, making unbiased estimators of dispersion difficult [244]. To resolve the lack of power, differential expression tools pool information across genes as per-gene variances in a given experiment are assumed to be similar [245]. Therefore, a single dispersion or fitted dispersion is estimated across all genes based on the assumption they share identical or similar variance. A second challenge is that of heteroskedasticity when lower counts tend to have higher variance and if left unaccounted, will have greater fold change between conditions and greater dispersion relative to the mean count. DESeq, DESeq2 and edgeR correct for heteroskedasticity by estimating dispersion as a function of the mean read count for each gene (and

therefore dispersion becomes $\alpha(\mu)$ instead of a per-sample scalar) [238]. The trended dispersion is estimated using a weighted conditional likelihood [232], [233]. In DESeq2, shrinkage estimators are used for dispersion and fold change, and therefore reduce variance for lower expressed genes at the expense of bias [246].

GLM methods can be applied to estimate the β coefficients in equation (3), which are interpreted as fold change (the relative contribution of condition β on the observed mean count μ), and statistical tests are applied to estimate whether β is significantly different from null (no difference). GLMs as in equation (3) can be additionally extended to account for multiple class memberships or interactions. edgeR and DESeq [238] both use an inference test similar to Fisher's exact test adapted for over dispersed data for pair-wise comparisons, while DESeq2 uses a Wald test [239]. edgeR, DESeq and DESeq2 use a likelihood ratio test to infer significance when applied to GLMs [238]-[240], [247].

An additional challenge differential expression tools face is to properly account for multiple testing correction as typically 10^4 - 10^5 simultaneous statistical inference tests are applied to the experiment [28]. This results in high false discoveries for a given p-value cutoff (i.e., comparing the observed value to the null distribution) and need to be adjusted. The simplest correction is the Bonferroni adjustment in which all p-values are divided by the total number of tests, and so the significance level α becomes α/n , adjusting the family-wise error rate [248]. An alternative approach is to use false discovery rate (FDR) estimation in which the proportion of null values are compared to the proportion of observed values at a given threshold cutoff [30].

In tests evaluating edgeR and DESeq, both methods performed equally well with regard to normalization, differential expression detection accuracy and low number of false positives, but had poor signal-to-noise ratio versus p-value correlation for genes detected in only one condition [249]. In a separate evaluation, edgeR and DESeq2 performed equally well regarding power, null hypothesis p-value uniformity and FDR control in two metagenomics evaluation datasets [250].

Although differential expression tools are invaluable to transcriptomics research, most differential expression tools are limited. First, they do not account for subject-specific effects or mixed effects. Additionally, tools without a linear model do not account for more than two pair-wise class comparisons or interactions among terms. Finally, conventional differential expression methods are not designed for time-series studies, continual or ordinal response variables, or mixed input variable types.

Machine learning approaches to identify relevant genes

Machine learning algorithms, broadly divided into supervised and unsupervised methods, are an alternative approach to classify, predict and identify patterns in RNA-seq data. In the context of RNA-seq supervised learning, gene products and other metadata (collectively known as “features” or “predictors”) are used to predict or classify an outcome (collectively known as a “response variable”, e.g., cancer vs. normal or treatment vs. control). Examples of supervised learning algorithms include Random Forest and Naïve Bayes. Conversely, in unsupervised learning, the goal is to uncover any underlying structure(s) of the data. Examples of

unsupervised learning algorithms include hierarchical clustering and k-means clustering.

Supervised machine learning uses features with known response variables to train a model, at which point data with unknown response variables can be classified or regressed using the trained model and new data. Model performance is assessed using a test set withheld during model training and where the correct response variable is known. The model is trained using the training set, and then new data predictions are compared against the true value to assess model accuracy (in classification, the proportion of true calls relative to all training data) and conversely, the error (the rate of misclassification). In regression, the mean squared error is often used to assess overall error of predicted from actual response:

$$\frac{1}{N} \sum_{n=1}^N (y_n - f(X_n))^2$$

Where y is the known value of sample n of N total samples and $f(X)$ is the predicted value of sample n . Furthermore, some algorithms implement a model assessment subroutine within the training phase where the training data is iteratively split into training and testing: k-fold cross validation or leave-one-out validation. In k-fold cross validation, the data is randomly subset into k sets (e.g., 10), then $k-1$ sets are used as the training set and model assessment is performed on the held-out set. This is repeated for all k -folds, and the overall model performance is aggregated across all folds [251]. In leave-one-out validation, a single sample is held out as the test sample [251]. Finally, to optimize model performance, parameters that govern the machine learning algorithm can be tuned independently by computing model

accuracy as a function of a range of parameter values and selecting the value that gives the best performance.

Random Forest is a supervised learning approach based on an ensemble of Classification and Regression Trees (CART). A CART is created by recursively partitioning the training set data according to the feature that maximizes the splitting criterion, where each split is known as a node [252]. The feature space is continually split until nodes are homogeneous. Thus, the root node contains the feature that maximally partitions the data, while all daughter nodes contain maximal splits conditioned on the criteria splitting the superseding node. Terminal nodes are assigned class labels based on the partitioning rules and new data can be therefore classified or regressed based on trained rules at each node [253], [254]. Random Forest implements the CART algorithm by generating an ensemble of K trees with two levels of bootstrapped randomization. The first level only considers a random subset of input features to construct each tree known as ‘bagging’, while the second level considers a random subset of samples to construct each tree in the forest [255]. The data that is not used to build a tree is termed ‘Out of Bag’ (OOB) and can be used as a test set to assess OOB error. Thus, the Random Forest procedure is:

- Bootstrap data
- Grow N trees using CART
 - For each node in a tree
 - Select a random subset of variables
 - Find best variable and cut point
- Determine OOB error using non-bootstrapped data
- Classify/regress new data using majority vote of N trees

The curse of dimensionality is reduced by training numerous decision trees on subsets of data, [256] and the effect is a more accurate classifier with high predictive value [256] over CART. Random Forest better accounts for variable interactions [256] as the random split of data ensures low correlation between trees. Unlike some other machine learning algorithms, Random Forest can perform regression. When compared to Support Vector Machine, Random Forest approaches showed 10% increased accuracy, which was attributed to the model and combined features [257].

Random Forest algorithms rank feature importance based on the feature's contribution to node purity or OOB prediction error, making them ideal for selecting a feature subset. Importance metrics are defined by a feature's total decrease in Gini impurity across all trees, or by the permutation-derived misclassification error (classification problems), mean squared error (regression) or sum of squared residuals (regression) [258], [259]. Permutations can additionally be used to generate a null distribution of randomized class labels to compare to the observed metric to calculate a p-value for the observed metric [260]. The number of candidate predictors, size of the trees, size of terminal nodes and resampling scheme (i.e. bootstrap samples drawn either with or without replacement) can be optimized to minimize OOB error [252]. Additionally, uninformative features can be iteratively removed to minimize error [261], [262].

To illustrate their utility, the following examples demonstrate where Random Forests were used to identify associations between gene expression and biological outcomes. A Random Forest model was used to stratify cancer patient responders to FOLFOX cancer therapy based on cancer type and found to be 69.2% accurate as

assessed by an independent test set [263]. In another example, a Random Forest algorithm was used to identify 196 stress response genes in the bacterium *Bacillus subtilis* wild-type and mutant strains [264]. In a third study, miRNAs were identified and shown to be significantly associated with the fold change of mRNAs using local importance and permutation [265].

Concluding Remarks

Extracting RNA carries special considerations given the delicate nature of RNA. Additional challenges exist for small RNA sequencing or miRNA-qPCR. After sequenced reads are mapped to the genome using an alignment tool such as BWA or Tophat/Bowtie, experimental questions can be addressed using parametric or non-parametric methods which attempt to identify transcripts most associated with a given outcome. Parametric methods discussed here rely on the Negative binomial distribution to compute differential expression and include edgeR, DESeq and DESeq2. Conversely, machine learning approaches attempt to find patterns in data, and include Random Forest models which are especially advantageous in situations in which data does not fit the design of parametric tools. Ultimately, regardless of the approach, the goal is to identify interesting transcripts to address biological questions, which will need to be validated experimentally *in vitro* or *in vivo*.

Specific Aims

The vaginal microbiota of reproductive-aged women is a highly dynamic environment in which the host senses microbes and microbial products. This process requires the coordination of gene transcription regulation which to date has not been fully realized. The role of miRNAs in controlling cellular responses to *Lactobacillus* spp.-dominated (CST I, II, III and V) and *Lactobacillus*-depleted (CST IV) vaginal microbiota remains unknown.

Using daily collected samples coupled with *in vitro* cell culture experiments, this project aims to characterize miRNA and miRNA-associated mRNA regulatory expression pathways as a function of vaginal microbiota community differences to gain insight into the functional interaction between the host and the microbiota.

Specific Aim 1

Determine miRNA differential expression as a function of *Lactobacillus* spp. and non-*Lactobacillus* community state types using Nugent-BV as the outcome. This aim leverages daily collected human vaginal samples to determine the types and abundances of human miRNAs to characterize host regulatory mechanisms. A machine learning approach was applied to discover miRNAs and their targets that are controlled by various vaginal CSTs

Specific Aim 2

Identify regulatory target pathways by simultaneously measuring vaginal epithelial mRNA transcripts as a function of characteristic vaginal communities. This

aim will discover the pathways and genes that play roles in host functions and will broaden the understanding of implications found in Chapter 2.

Chapter 2 microRNA expression induced by vaginal microbiota controls host epithelial cell proliferation and susceptibility to *Chlamydia trachomatis*

Introduction

The human vaginal microbiota is comprised of at least five major bacterial Community State Types (CSTs), of which four are often dominated by species of *Lactobacillus* (CST I: *L. crispatus*; CST II: *L. gasseri*; CST III: *L. iners*, CST V: *L. jensenii*) while the fifth (CST IV) is characterized by a paucity of *Lactobacillus* spp. and a diverse, more homogenous set of strict and facultative anaerobes such as *Gardnerella vaginalis*, *Atopobium vaginae*, *Prevotella*, *Megasphaera* and *Mobiluncus* among others [6]. Microbiologically, CST IV is similar to the community state found in cases of bacterial vaginosis (BV) [9], [266]-[268]. BV is typically diagnosed in one of two ways. The Amsel's criteria, applied in a clinical setting, define Amsel-BV as presenting with three of the four following signs: vaginal pH greater than 4.5, fishy odor upon exposure of a vaginal smear to potassium hydroxide, presence of clue cells upon microscopic examination, and an abnormal discharge [13]. The Nugent scoring system, used in research settings, grades the relative presence of *Lactobacillus* versus *Gardnerella* and *Mobiluncus* morphotypes on Gram stained vaginal smears, thus defining Nugent-BV [11], [12]. Nugent scores range from 0 to 10, with a score 7 to 10 considered BV, 4 to 6 intermediate, and 0 to 3 normal. Not surprisingly, high Nugent scores are often associated with CST IV, and thus encompass both symptomatic and asymptomatic BV states. As such, BV (Amsel- or Nugent-BV) is often described as a state of dysbiosis, which has been shown to be a risk factor for the acquisition of sexually transmitted infections (STIs), including HIV, *Neisseria*

gonorrhoeae, and *Chlamydia trachomatis* [28]-[30], [269]-[271]. By definition, BV does not elicit a consistent and pronounced immune response, and BV-specific cytokine and chemokine profiles are varied [47]. Microbes associated with BV may actively suppress or induce a low-grade host response [272], or subclinical immune response [273]. Alternatively, the host may be tolerant to BV microbes [274]. Thus, the host response to different types of vaginal microbiota is poorly understood at the molecular level. Characterizing the molecular mechanisms elicited by CST-IV and BV may help better understand and manage the condition, and concurrently develop strategies to reduce STI susceptibility and other associated co-morbidities.

Human microRNAs (miRNAs) are ~20-mer oligonucleotides that have been shown to regulate a myriad of functions via translational inhibition [144]. Their role in controlling cellular responses to *Lactobacillus*-dominated (CST I, II, III and V) and *Lactobacillus*-depleted (CST IV) vaginal microbiota remains unknown even though they may constitute an underappreciated host response to changes in the vaginal microbiota. This aim leverages daily collected human vaginal samples to determine the types and abundances of human miRNAs and characterize host regulatory mechanisms that drive the response to vaginal microbiota communities associated with *Lactobacillus* spp.-dominated states and Nugent-BV. A machine learning approach was used to discover miRNAs and their targets that are controlled by various vaginal CSTs. From this effort, miR-193b was identified as a candidate effector miRNA and its role in suppressing vaginal epithelial cell proliferation in *Lactobacillus*-dominated communities was confirmed *in vitro*. Furthermore, because Nugent-BV is a risk factor for chlamydial infection, epithelial cell proliferation was

demonstrated to be required for efficient *C. trachomatis* infection. These findings are critical and contribute to the fundamental understanding of the role of the vaginal microbiota in susceptibility or control of STI. Further, this aim has broad clinical implications for the management of women's health and could be leveraged to develop novel preventive therapies for STI in general.

Methods

Vaginal swab sample collection and metadata

The samples used in this aim were collected as part of a previously assembled cohort of 135 reproductive-age non-pregnant women at the University of Alabama at Birmingham in which participants were instructed to sample their vagina daily for 10 weeks [19]. The clinical study protocol was approved by the Institutional Review Board of the University of Alabama at Birmingham and the University of Maryland School of Medicine. Written informed consent was appropriately obtained from all participants. Swabs were stored at -80°C in 2 ml of Amies Transport Media/RNA later (50%/50% v/v), a solution formulated to stabilize RNA (QIAGEN). In that study, subjects answered daily questionnaires on sexual activity, menstruation and symptomatology among others and provided vaginal swab smears on glass slides for subsequent Gram staining and Nugent scoring by a trained technician. Estimated time in menstrual cycle as a Random Forest predictor were normalized using previously described methods in which the time between self-reported menstrual periods is scaled to 28 days [7] and then categorized into either the menses phase (day 1-5), proliferative phase (day 6-13) or secretory phase (day 14-28). Subjects also

underwent a gynecological exam at enrollment, week 5 and week 10, during which a clinician used Amsel criteria for the diagnosis of BV. The composition of vaginal microbiota was assessed by metataxonomics using amplicon sequencing of the 16S rRNA gene V3-V4 hypervariable regions [7], [275]. Taxonomy was assigned to each read using a novel and fast Markov Chain taxonomic classifier (available at <http://ravel-lab.org/pecan>) and taxa frequencies normalized to total per-sample read counts. Community State Types (CSTs) were identified by calculating the Jensen-Shannon divergence among samples, followed by hierarchical clustering with complete linkage [6]. Subjects included in NBV, PBV and TBV groups were selected per criteria described above.

Total RNA extraction

Total RNA extraction from vaginal swabs was performed in random order on selected samples across multiple days and subjects to minimize batch effects. The MasterPure™ Complete DNA and RNA Purification Kit (Epicentre, # MCR85102) was used to extract total RNA from 250 µl aliquot of swab storage buffer containing RNAlater or confluent adherent VK2 epithelial cells following the manufacturer's recommendations for cell samples. Swab material or cells were centrifuged for 10 minutes at 13,000 x g, supernatant removed, and lysed with 300 µl Lysis buffer containing 50 mg Proteinase K. The pellet was incubated for 15 minutes at 65°C with 10-second vortexing every 5 minutes, before placing on ice for 5 minutes. Following this step, 175 µl Protein Precipitation buffer was added and the mixture vortexed for 10 seconds, and centrifuged for 10 minutes at 13,000 x g. The supernatant containing total nucleic acid was added to 500 µl isopropanol and mixed by inverting the tube 40

times to precipitate nucleic acid. Nucleic acids were pelleted by centrifugation for 20 minutes at 13,000 x g and washed twice with 70% ethanol, left to air-dry for 5-15 minutes. The pellet was resuspended in 10 μ l nuclease-free water. VK2 epithelial DNA was then removed by adding 1 μ l TURBO DNA-free DNase and incubated for 30 minutes at 37°C followed by an additional treatment with 1 μ l of DNase for another 30 minutes (Life Technologies, # AM1907). DNase was inactivated by adding 2 μ l DNase Inactivation Buffer and incubated for 5 minutes at room temperature. The mixture was centrifuged at 13,000 x g for 3 minutes and the supernatant containing total RNA transferred to a new tube. The RNA quality and quantity was measured using 1 μ l of RNA solution with a 2200 Agilent TapeStation and RNA screen tape (Agilent # 5067-5576). DNA was then removed by adding 1 μ l TURBO DNA-free DNase and incubated for 30 minutes at 37°C followed by an additional treatment with 1 μ l of DNase for another 30 minutes (Life Technologies, # AM1907). DNase was inactivated by adding 2 μ l DNase Inactivation Buffer and incubated for 5 minutes at room temperature. The mixture was centrifuged at 13,000 x g for 3 minutes and the supernatant containing total RNA transferred to a new tube. The RNA quality and quantity was measured using 1 μ l of RNA solution with a 2200 Agilent TapeStation and RNA screen tape (Agilent # 5067-5576). Total RNA was stored up to 1 week at -80°C until further use. All samples yielded at least 20 ng total RNA. Although RIN values were of sub-optimal quality for full-length transcript RNA-seq, miRNAs were generally intact and adequate for small RNA-seq as evidenced by a lack of correlation between miRNA reads and RINe (Figure 2.2).

Small RNA sequencing library construction

All small RNA-seq library preparations were randomized across samples from multiple days and subjects to minimize batch effects and were carried out using the TruSeq Small RNA kit per manufacturer's recommendations (Illumina, # RS-200-0012), with CleanTag Ligation from TriLink Biotechnologies' modifications (TriLink Biotechnologies, # L-3203) at 1:3 5' and 3' adaptor dilutions and 15 cycles for library enrichment. Briefly, both the 3' CleanTag Adaptor and 5' Adaptor were diluted 1:3 in nuclease-free water to accommodate 50-100 ng total RNA input. The RNA template was denatured for 2 minutes at 70°C, then 1 µl diluted 3' adaptor, 1 µl RNase Inhibitor, 1 µl Enzyme 1 and 5 µl Buffer 1 were added to the template, mixed, and incubated at 28°C for 1 hour followed by incubation at 65°C for 20 minutes. Following this step, 4 µl nuclease-free water, 1 µl Buffer 2, 1 µl RNase Inhibitor, and 2 µl Enzyme 2, was added to the RNA template and the 3' adaptor mixture. The diluted 5' adaptor was denatured for 2 minutes at 70°C, then 2 µl was added to the mixture and incubated at 28°C for 1 h followed by an incubation at 65°C for 20 minutes. The tagged library underwent reverse transcription by adding 2 µl RT primer (TruSeq kit), 1.92 µl RNase-free water, 5.76 µl RT buffer (SuperScript II/Life Tech), 1.44 µl dNTPs, 2.88 µl 0.1 mM DTT, 1 µl RNase Inhibitor, 1 µl superScript II (Life Technologies), and then incubated at 50°C for 1 hour. The cDNA was enriched by PCR by adding 40 µl 2X Phusion High Fidelity Taq Polymerase Mastermix (ThermoFisher), and 2 µl each of the universal forward primer and a sample-specific index, then PCR amplified using the following conditions: 98°C for 30 seconds, [15 cycles of 98°C for 10 seconds, 60°C for 30 seconds, 72°C for 15 seconds] and a final

extension at 72°C for 10 minutes. The enriched libraries were purified using Agencourt AMPure XP beads by adding 80 µl beads to the 80 µl reaction volume and incubating for 10 minutes to bind DNA. The beads were magnetized for 4 minutes, then the supernatant containing the library was transferred to a new tube where 144 µl beads were added and incubated for 10 minutes to bind DNA. The beads were magnetized again for 4 minutes, the supernatant discarded, and the beads washed twice with 500 µl 70% ethanol. After the wash, the beads were left to air-dry before resuspending in 17 µl nuclease-free water for 2 minutes. The solution was re-magnetized and 15 µl was transferred as the small RNA-seq library. Libraries were validated on the LabChip GX (PerkinElmer) (approximately 145 bp size) before cluster generation and sequencing.

Small RNA sequencing, quality control and read mapping

Small RNA-seq libraries were sequenced on a HiSeq 4000 with either 75 bp single-end (SE) or 150 bp paired-end (PE) reads, at about 20-40 million reads per library (approximately 10% sample per lane). Reads from the R1 fastq file (150 PE) or R2 fastq file (75 SE) were trimmed using Trimmomatic-0.33 with the following parameters: LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15 and MINLEN:16, then visualized for quality using fastqc (version 0.10.0). Reads were aligned with bwa (version 0.5.9-r16) using 2 as the maximum number of mismatches between reference and read to the following series of references, in which all unaligned reads were aligned to the next in series: first, all tRNA from the GtRNADB (<http://lowelab.ucsc.edu/GtRNADB/>) [276], then human rRNA (hum5SrDNA and humRibosomal each a part of Illumina's iGenomes available at

ftp://illumina.com/Homo_sapiens/UCSC/hg19/Homo_sapiens_UCSC_hg19.tar.gz, downloaded August 17, 2015), *G. vaginalis* ATTC 14019 (NCBI reference NC_014644.1) [277], *L. iners* ATCC 55195 (NCBI reference NZ_GL622333.1), and finally human hg19 (also downloaded from Illumina's iGenomes as above). Reads aligning to human miRNA regions were annotated using HTSeq (version 0.5.3p3, Python version 2.7) and annotations from primary transcript miRNAs from the miRBase v20 (GTF version 3, GRCh37.p5, NCBI Assembly GCA 000001405.6). Samples with low miRNA reads were removed from further analysis, which were defined as less than 125 reads/human miRNA [$125 * 1869$ (total number of annotated miRNA in GTF) = 233,625 total reads/sample]. An additional QC was performed using PCA plots to identify batch or subject-specific effects (prcomp, version 3.3.1, ggbiplot, version 0.55, gPCA version 1.0 [278]). Reads surviving trimming and QC were normalized using edgeR's (3.6.8) calcNormFactors function [240], which uses a Trimmed-Mean of M values approach as described in [242].

microRNA qPCR

RNA samples were converted into polyadenylated cDNA for qPCR analysis by mixing 7 μ l (50-100 ng) total RNA in water with QIAGEN miScript II RT (QIAGEN, # 218160) reagents: 2 μ l Reverse Transcriptase Mix, 2 μ l 10X Nucleic acids Mix, 4 μ l 5X miScript HiSpec Buffer and 5 μ l nuclease-free water. The reaction mixture was incubated at 37°C for 1 hour, then heated to 95°C for 5 minutes to stop the reaction per the manufacturer's protocol [221], [279]. The resulting cDNA was diluted 1:10 in water before use in the subsequent qPCR assay.

qPCR reactions were performed using 1 µl diluted cDNA template, and the miScript SYBR Green PCR Kit by mixing 5 µl 1X SYBR Green Mastermix and 1 µl each of Universal and miRNA-specific primers in a 10 µl reaction volume (QIAGEN, # 218073 and # 218160, miR-specific primers: MS00031549 (miR-193b-3p, 5'AACUGGCCCUCAAAGUCCCGCU) and MS00033740 (RNU6-2-11)). qPCR was carried out on a 7900HT thermocycler at the following cycle conditions: 95°C for 15 minutes, then 40 cycles of: 94°C for 15 seconds, 55°C for 30 seconds and 70°C for 30 seconds. Ct values were calculated using Applied Biosystems SDS software and used to compare targets within and between samples by normalizing the within-sample mean miR-193b Ct to the within-sample mean Ct for RNU6-2-11 (Δ Ct) [280], [281]. For small RNA-seq validation, a linear mixed-effect model (nlme package version 3.1-128) was used to compute the probability that the effect due to PBV relative to NBV Δ Ct is not zero, while controlling for subject specific (random) effects. For all other qPCR experiments, the $\Delta\Delta$ Ct method was used to compute the Δ Ct between miR-193b and RNU-6 within a sample, and then the Δ Ct between *Lactobacillus* spp. BCS and cell culture medium or *G. vaginalis* BCS (non-*Lactobacillus* spp.) [281]. The $\Delta\Delta$ Ct standard deviation was computed using $\sigma_{\Delta\Delta Ct} = (\sigma_{\Delta Ct \text{ Lactobacillus spp.}}^2 + \sigma_{\Delta Ct \text{ non-Lactobacillus spp.}}^2)^{1/2}$, where σ is the standard deviation. A two-tailed t-test was applied to each $\Delta\Delta$ Ct, with $p < 0.05$ considered significant.

Identification of BV-associated taxa and BV-associated miRNAs using Random Forest

Random forest models were applied to select taxa or miRNAs that best predict non-BV from BV status defined either as Amsel-BV or Nugent-BV by utilizing a

combination of the R software packages `rfPermute` (version 2.0.1, [282]), `randomForest` (version 4.6-12, [259]) and custom subroutines (available in the R scripts ‘`rfSubjectSpecific.R`’ and ‘`Smith_et_al_AnalysisScript.Rmd`’ included in Appendix 2). All models were trained using 70% of available data (training set) while model performance was assessed using the remaining 30% of the held-out data (testing set). Because Amsel diagnosis data is not available for most of the samples in the study, another Random Forest model was trained on the larger sample cohort to classify a sample’s Amsel diagnosis based on metataxonomic data (expressed as taxa relative abundance, available in SRA under project PRJNA208535) and other metadata (normalized mensuration as described above, vaginal pH, vaginal intercourse, oral sex, anal sex, finger penetration, sex toy use, hormonal birth control use, BV symptoms (vaginal odor and discharge) and non-BV symptoms (vaginal itching, burning and irritation)). The trained model was then used to assign proxy-Amsel-BV diagnoses to each of the samples in the study given taxa relative abundance and metadata (which was available for 83 samples). Each sample’s Nugent score (discrete 0-10, inclusive) or proxy-Amsel diagnosis (classification of “NBV” or “PBV”) were used as the response variable to determine the most important miRNAs or metadata that predicted BV state (models termed Nugent-RF or proxy-Amsel-RF, respectively). The Nugent score was used in a regressive capacity as it best captures the ordinal relationship reflective of differences in bacterial morphotypes on Gram stained slides. Because of the continuous nature of the Nugent score values in the model prediction, accuracy assessment was calculated using mean error:

$$\frac{1}{N} \sum_{n=1}^N |predicted_n - actual_n|$$
, where N is the number of test set samples. Accuracy

for the Amsel Random Forest and proxy-Amsel scoring were assessed using the Amsel-BV diagnosis or proxy-Amsel-BV diagnosis from the test set, respectively.

In both Nugent-RF and proxy-Amsel-RF models, the \log_2 transformed normalized miRNA counts in addition to metadata (normalized mensuration as described above, vaginal intercourse, oral sex, anal sex, finger penetration, sex toy usage, hormonal birth control usage, BV symptoms (vaginal odor and discharge) and non-BV symptoms (vaginal itching, burning and irritation)) were used as predictors. To increase the confidence of feature calls, miRNAs with zero counts in any sample were excluded, as zero miRNA counts could be due to under sampling. Each model underwent 10-fold cross-validation, with 500 permutations to determine the null distribution for p-value calculation. Default parameters for ntree (500) and mtry ($\sqrt{1558}=39$ for proxy-Amsel-RF, $1558/3=519$ for Nugent-RF) were used. The algorithm accounted for non-independent samples that originated from the same subject by evenly splitting each cross-fold iteration among subjects using an in-house script (available in Appendix 2 as 'rfSubjectSpecific.R'). Importance metrics and p-values were calculated based on rfPermute and randomForest R packages [259], [282]. The importance metrics for the proxy-Amsel-RF model were mean decrease in accuracy and mean decrease in Gini coefficient while the importance metrics for the Nugent-RF model were increase in mean squared error and increase in node purity [256], [283]. Statistically significant features were defined as features with p-value < 0.05 for any importance metric within a model result.

Experimentally validated miRNA targets were identified using the “strong evidences” list from miRTarBase, Release 6.0 (Sept 15, 2015) [284]. The Gene

Ontology DIRECT process terms from DAVID (release 6.8, May 2016, [285]) were mapped to experimentally validated miRNA targets. The proportion of targets for each GO DIRECT term were computed for each miRNA.

Scratch assay using Bacterial Culture Supernatant (BCS)

VK2 epithelial cells (ATCC CRL-2616, cell line authentication report in Appendix 3) were cultured at 37°C in 5% CO₂ then seeded at 7.5x10⁴ cells/well and grown to confluence in VK2 complete medium [Keratinocyte SFM, ThermoFisher # 17005042, with bovine pituitary extract (0.05 mg/ml), epidermal growth factor (0.1 ng/ml) and CaCl₂ (0.4 mM)]. VK2 cells were starved using base medium (Keratinocyte SFM only) for 24 hours before making scratches in the monolayer with a 1 ml sterile pipette tip. BCSs were created by seeding 1x10⁷ bacteria/mL of either *L. crispatus* (ATCC 33197), *L. jensenii* (ATCC 25258), *L. iners* (ATCC 55195) or *G. vaginalis* (ATCC 14018) in 10 mL culture media (NYC-III for *L. crispatus*, *L. jensenii* and *L. iners*, TSB for *G. vaginalis*), grown anaerobically for 48 hours, centrifuged at 3,000 x g for 10 minutes, sterile filtered (2.2µm filter) and stored at -20°C. Lactic acid conditioned media were prepared using 0.1% or 0.06% of either D(-)-lactic acid, (Sigma, #L0625), or L(+)-lactic acid (Sigma, #L1750) in VK2 cell culture medium as the 0.1% concentrations approximated the concentration found in 20% BCS (as quantified by D/L lactic acid assay kit, R-Biopharm # 11112821035). pH buffered 7.66 1% DL-lactic acid was prepared using approximately 10% (v/v) 1N hydrochloric acid into 1% racemic DL-lactic acid. BCSs were diluted to 20% (v/v) using complete VK2 cell culture medium and added to VK2 cells for a period of up to 24 hours. Images were taken at 100X under phase contrast using a Zeiss Primovert

microscope at 0 and 13 hours post BCS exposure. ImageJ software (version 1.50i, [286]) was used to quantify the proportion of cells occupying the scratched area relative to time 0 h in three fields per well. After imaging for cell proliferation, 300 μ l RNAlater was added to wells, cells were mechanically detached from plate, and RNA was extracted from cells using the total RNA extraction protocol described above. Replicate wells were separately stained for viability/cytotoxicity per manufacturer's instructions (4 μ M Calcein AM and 4 μ M Ethidium homodimer III in PBS, Biotium, # 30002) and imaged using a Nikon TE2000-E2 fluorescence microscope at 100X using a FITC (viable) or a TRITC (cytotoxic) filter. To monitor cells for active DNA synthesis, EdU (5-ethynyl-2'-deoxyuridine) was added to cell culture medium at 10 μ M with cells grown on coverslips at the 0 h time point. Cells were fixed using 100% methanol at the end time point, washed with 3% BSA, permeabilized using 0.5% Triton® X-100 in PBS, washed with 3% BSA and stained for 15 minutes with Alexa 488 mixture per manufacturer's instructions (ThermoFisher, # C10337). Cell nuclei were stained for 15 minutes using Hoechst 33342 at a 1:1,000 dilution in PBS (ThermoFisher), then rinsed once in PBS before being imaged at 100X using a Zeiss Axio Imager Z1 fluorescence microscope and the GFP (EdU) or Dapi (Hoechst) filters. The amount of DNA synthesis was calculated using CellProfiler (version 2.2.0 rev 9969f42, [287]) by counting the number of green nuclei (EdU stained) relative to blue nuclei (DAPI stained) in five fields per duplicate coverslip. A two-tailed t-test was applied to test whether differences between means of each scratch assay and EdU experimental conditions were equal to 0, with $p < 0.05$ considered significant (See Appendix 2 for script).

Cell Cyclin D1 (CCND1) Western blot

VK2 epithelial cells were grown and exposed to respective BCSs as above for 13 hours, then media removed and washed once with cold PBS. Total protein was extracted using 500 μ l RIPA buffer (Cell Signaling, #9806S) for 5 minutes on ice. Cell debris was pelleted by centrifugation at 13,000 \times g for 15 minutes and supernatant containing total cell lysate stored at -80°C until further use. Protein quantification was determined using the Bradford method assay (Bio-Rad # 500-0205). Tris-glycine precast gels, 4-15% (Bio-Rad # 456-1086), were loaded with 20 μ g total protein per well and run for 35 minutes at 140 V before transferring to a PVDF membrane at 20V for 20 minutes. Membranes were blocked using Odyssey PBS blocking buffer (Li-Cor part # 927-40000) for 1 hour. Primary antibodies for purified mouse anti-human Cyclin D1, clone G124-326 (1:300, BD Pharmingen, # 554180, lot # 5357909) and rabbit anti β -actin (1:5,000, Abcam, # 8227, lot GR297401-1) were incubated with blocking buffer and 0.2% Tween-20 for 1 hour. Membranes were washed four times with PBS-0.1% Tween-20 and then incubated with 1:15,000 secondary antibodies (Li-Cor, IRDye 680RD goat anti-rabbit, lot # C60920-09 and IRDye 800CW goat anti-mouse, lot # C61012-06) before a final wash and imaging using Odyssey[®] CLx Imager and Image Studio software (version 5.2). Protein band intensities were measured using Image Studio software, then CCND1 intensity values were normalized to β -actin loading control value.

C. trachomatis infection and inhibition of cell proliferation

C. trachomatis serovar L2 (L2) was propagated in HeLa monolayers as described previously [288]. Briefly, L2 was cultivated in 100 mm² tissue culture dishes in Dulbecco's modified Eagle's medium (DMEM) (Mediatechm) supplemented with 10% FBS at 37°C, 5% CO₂. Monolayers were gently rocked for 2 hours, rinsed, fresh media added and the infection allowed to progress for 48 hours. Elementary Bodies were harvested, (stock 9.6 x 10⁷ IFU/ml), and stored at -80°C. L2 was used directly from -80°C stock for all experiments.

A2EN human cervical epithelial cells (kindly provided by Dr. Quayle, cell line authentication report in Appendix 3) [289] were cultured at 37°C in 5% CO₂ in EpiLife Media supplemented with EpiLife Defined Growth Supplement (EDGS) and L-glutamine (ThermoFisher , GIBCO M-EPI-500-CA, GIBCO S-012-5 and #25030156). A2EN cells were seeded at 1x10⁵ on coverslips and grown overnight, then rinsed with PBS and exposed to starvation medium (EpiLife medium without EDGS supplement) for 18 hours. Coverslips were exposed to 300µl of either complete medium (control), CDK4 inhibitor CAS 546102-60-7 (400nM, Millipore #219476) or Fascaplysin (350nM, Millipore #341251) in complete medium. Concurrently, *C. trachomatis* serovar L2 was added at MOI 2 to each coverslip and rocked for 2 hours at room temperature. A2EN cells were rinsed with PBS and exposed to complete medium or inhibitors in the presence of EdU (1nM) as described above for an additional 22 hours at 37°C in 5% CO₂, then washed with 500 µl PBS and fixed with 95% methanol for 10 minutes (Sigma #M3641). Cells were incubated in the dark for 1 hour in 1:10 MicroTrak (Trinity Biotech #8H019UL) to stain *C.*

trachomatis cells. EdU (Alexa Fluor 555, ThermoFisher #C10638) and cell viability staining and imaging were performed as outlined above. The proportion of infected A2EN cells and new DNA synthesis were determined using manual counting and CellProfiler as above, respectively. A two-tailed t-test was applied to test whether differences between means of each EdU or infection experimental conditions were equal to 0, with $p < 0.05$ considered significant. A linear model was fitted to mean percent cells stained with EdU versus the mean proportion of infected A2EN cells using R lm package (version 3.2.3) (See Appendix 2 for script).

Results

Subject vaginal microbiota profile selection and small RNA-seq

From vaginal microbiota profiles previously characterized by metataxonomics analysis (16S rRNA gene sequencing) of samples collected daily over 10 weeks by 135 reproductive-aged women (parent cohort) [19], subsets of samples were selected (Figure 2.1) that were characterized by: 1) CSTs that were persistently dominated by *Lactobacillus* spp. with few or no reported vaginal symptoms and no diagnosis of BV based on Amsel criteria for all three clinical visits during the study period (“Negative BV-state”, NBV), 2) persistent Nugent-BV-associated CSTs (“Persistent BV-associated state”, PBV) that were sometimes accompanied by vaginal symptoms, and the subject was Amsel-BV positive for at least one of the three clinical visits during the study period, and 3) at least one transition between *Lactobacillus* spp. dominance and Nugent-BV associated CSTs (“Transition NBV-PBV”, TBV). Neither daily behavior (vaginal intercourse or oral sex) nor menstruation were consistently

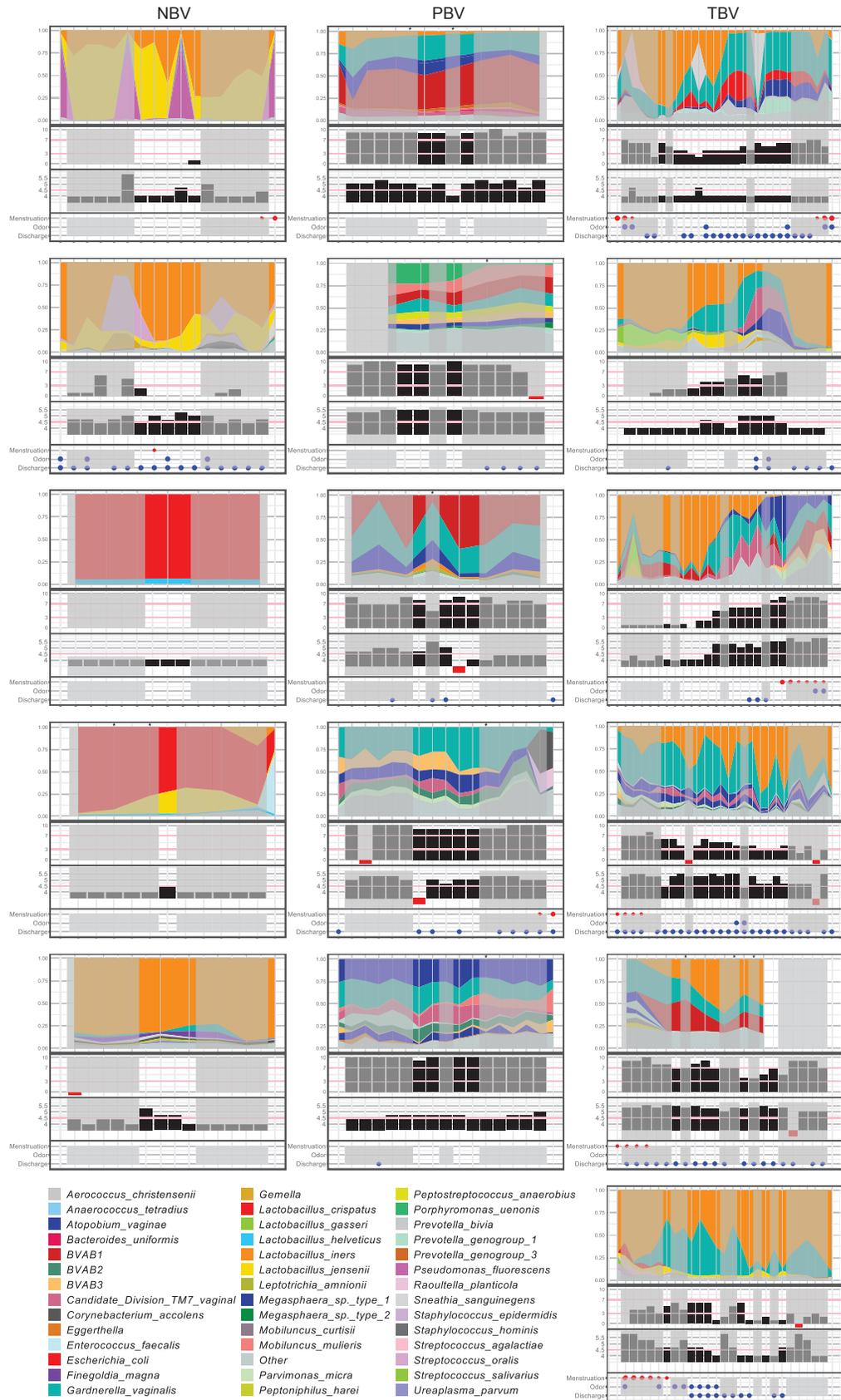


Figure 2.1 Longitudinal profiles of vaginal microbiota for each subject in the study.

From top to bottom for each subject, panels show the profile of relative bacterial abundance, Nugent score (0 to 10), pH (4 to 5.5) and metadata including reported menstruation (red dots), odor (blue dots, middle row) and discharge (blue dots, lower row). Greyed vertical blocks within the profiles represent days in which a vaginal swab was collected but not analyzed while non-greyed blocks are samples used in the study. An asterisk above a sample indicates that it was dropped after failing QC. Note that not all longitudinal windows are the same length as TBV subjects were generally sampled more frequently to capture changes in vaginal microbiota composition. Red bars in Nugent score and pH bar plots denote missing data.

associated with the differences in vaginal microbiota composition and structure of vaginal communities of participants as previously demonstrated [7], [19].

Five subjects each were selected from the NBV and PBV groups and six subjects were chosen from the TBV group. From these, vaginal samples (1-16 per subject) were sequenced using small RNA-seq (Appendix 4). These sequences and their abundances were used in a comparative analysis of NBV and PBV states identified in the three groups (Figure 2.1). Of the 113 samples sequenced, 13 failed QC due to low miRNA read counts, leaving 100 samples from 16 subjects from the NBV, PBV and TBV groups that were used in the final post-QC analysis (Table 2.1, Appendix 5). The median RNA Integrity Number (RINe) score of all samples was 6.6, with 60% of the samples having a RINe greater than 7 (Appendix 6). There was no relationship between the number or percentage of human genome (version 19, hg19) aligned miRNA reads and the RINe values (Figure 2.2), so that a RNA quality threshold was not imposed for inclusion in the study. PCA plots before and after sample removal and after normalization do not support batch effects (guided PCA p-values 0.99, 0.26, and 0.12, respectively), or subject-specific effects (guided PCA p-values 0.29, 0.48, and 0.29, respectively, Figure 2.2). The median (minimum-maximum) percentage of miRNA reads relative to all hg19 mapped and total reads was 26.6% (3%-58.2%) and 4.2% (0.5%-21%), respectively. The median number of

Table 2.1 Sample groups and number of samples pre- and post-sequencing QC, per subject			
BV Group	Subject ID	Number Samples before QC	Number Samples after QC
NBV	EM12	5	5
NBV	UAB007	5	5
NBV	UAB093	3	3
NBV	UAB102	3	1
NBV	UAB117	4	4
PBV	UAB008	5	3
PBV	UAB021	4	3
PBV	UAB022	5	4
PBV	UAB055	6	5
PBV	UAB116	5	4
TBV	UAB003	16	16
TBV	UAB005	6	5
TBV	UAB006	13	12
TBV	UAB015	14	14
TBV	UAB115	9	9
TBV	UAB121	10	7

post-QC hg19 mapped miRNA reads was 1,204,913 with a minimum of 239,758 reads and a maximum of 4,642,910 reads. Thus, despite a low proportion of miRNA reads, the estimated coverage ranged from 128X-2,484X across the entire miRNome database of 1,869 annotated miRNAs (Appendix 6).

Prediction of Amsel-BV diagnosis (proxy Amsel-BV) using metataxonomic data and metadata

An assessment of BV status using Amsel criteria was only done at three time points during the 10-week study (enrollment, 5-week and 10-week clinical visits, [19]), thus to identify miRNA associated with Amsel-BV, a model to predict Amsel diagnosis (proxy-Amsel-BV) using metataxonomic data and metadata was first

developed. This was accomplished by applying a Random Forest model trained with the metataxonomic data and metadata from 117 subjects of the parent cohort that included 281 samples for which both metataxonomic data and Amsel diagnosis was available (Amsel subset) (Appendix 7, Appendix 8 and Appendix 9). In the Amsel subset, asymptomatic or symptomatic BV diagnoses represented 25.3% (71/281) of the Amsel test performed (Appendix 7), a figure closely matching the reported prevalence of 29.2% for BV in similar populations [290]. The Amsel Random Forest model accuracy was tested using a hold-out set and found to be 94.9% accurate in correctly assigning NBV diagnosis and 72.0% accurate in correctly assigning PBV diagnosis. Multiple (34) important microbial (taxa and their relative abundance) or metadata features were predictive of the Amsel diagnosis: *Aerococcus*, *Aerococcus vaginalis*, *A. vaginae*, *Bacteroides coagulans*, *Bifidobacterium bifidum*, BVAB1, BVAB2, BVAB3, Candidate Division TM7 vaginal, *Dialister* sp. type 2, *Eggerthella*, *Eubacterium rectale*, *G. vaginalis*, *Gemella*, *L. crispatus*, *Lactobacillus helveticus*, *Leptotrichia amnionii*, *Megasphaera* sp. type 1, menstruation, *Mobiluncus*, *Mobiluncus mulieris*, *Parvimonas micra*, *Peptoniphilus lacrimalis*, vaginal pH, *Porphyromonas endodontalis*, *Porphyromonas* sp. type 1, *Porphyromonas uenonis*, *Prevotella* genogroups 1- 4, *Prevotella melaninogenica* and *Streptococcus salivarius* (Appendix 7). A proxy-Amsel-BV diagnosis was then assigned to each sample for which small RNA-seq was performed, using the Amsel Random Forest model within the NBV,

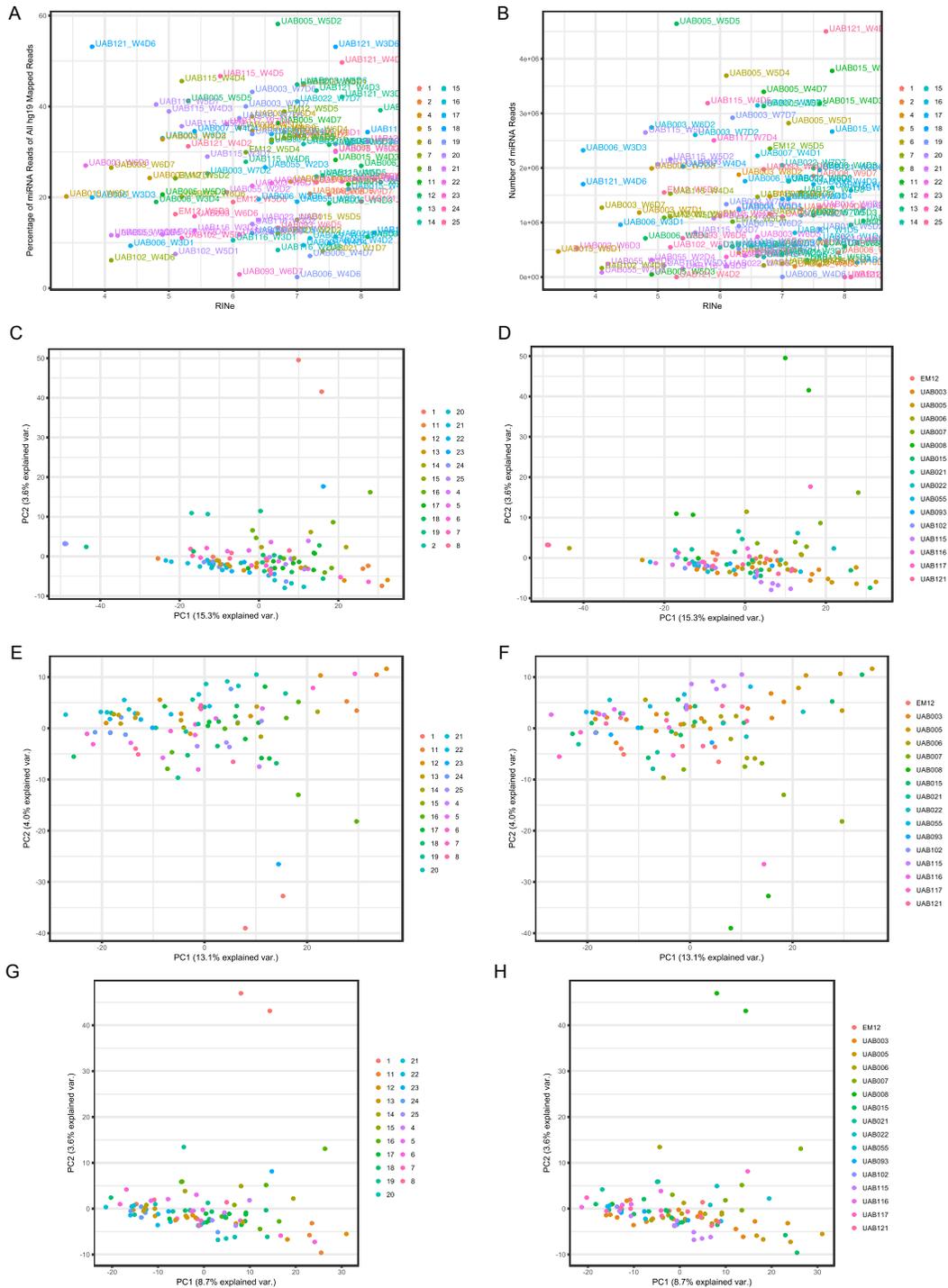


Figure 2.2 Quality Control (QC) figures of small RNA-seq samples.

(A and B) Relationship of total RNA quality (RINe, x-axis) versus miRNA mapped reads colored by sequencing batch as (A) a percentage of all hg19 mapped reads or (B) total number of miRNA reads. PCA plot of log₂ transformed miRNA read counts (C and D) before and (E and F) after low count read removal and (G and H) after normalization, colored by sequencing batch (left column) or subject (right column).

BV or TBV sub groupings, based on available metataxonomic relative abundances and behavioral metadata (Appendix 9)

Feature selection to identify miRNAs predictive of Nugent-BV and Proxy-Amsel-BV

Two different Random Forest models were built to identify miRNAs predictive of Nugent score (Nugent-RF model) or proxy-Amsel-BV diagnosis (proxy-Amsel-RF model). The Nugent-RF and proxy-Amsel-RF models used 178 predictors including 169 non-zero log₂ transformed miRNA read counts and 9 metadata variables as inputs to rank feature importance (Appendix 9). The accuracy of the proxy-Amsel-RF model classification was 82.3% for NBV and 80.3% for PBV and the Nugent-RF model on average correctly predicted the Nugent score within 2 values. There were 20 and 30 miRNAs with model importance permutation p-value < 0.05 identified using proxy-Amsel-RF and Nugent-RF, respectively (Figure 2.3, Table 2.2, Appendix 10). A total of 8 miRNAs were identified by both models: miR-193b, miR-182, miR-203b, miR-378a, miR-3607, miR-324, miR-500a, and miR-146a. The expression profiles of these 8 miRNAs as a function of CST, Nugent score and proxy-Amsel-BV prediction are shown in Figure 2.3.

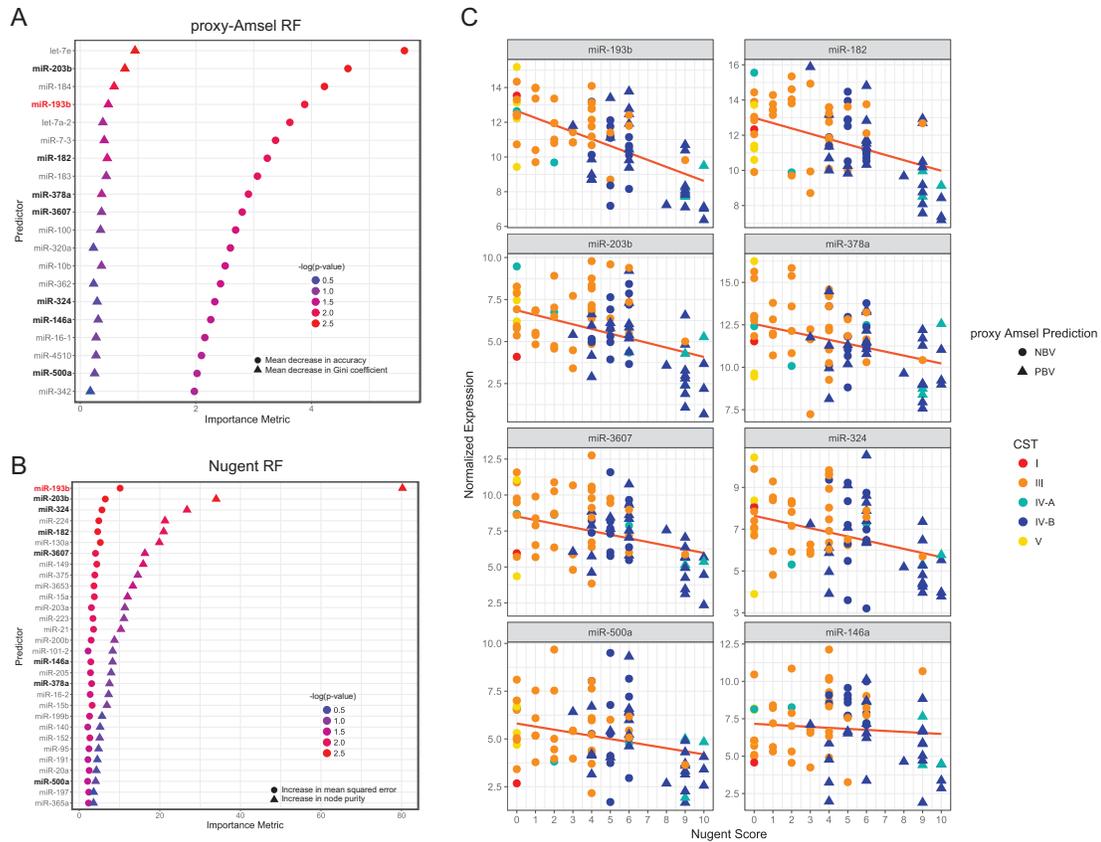


Figure 2.3 Random Forest variable importance ranking for proxy-Amsel-RF and Nugent-RF models and miRNA expression of the eight miRNAs identified by both models.

Random Forest model variable ranking using either (A) proxy-Amsel-BV diagnosis or (B) Nugent score. Features ranked by variable importance metric for mean decrease in accuracy (proxy-Amsel-BV) or increase in node purity (Nugent). The most significant features are listed for each model. Importance metrics are colored by $-\log_{10}$ p-value. (C) \log_2 normalized miRNA expression values are shown versus Nugent score. Circles indicate predicted PBV by proxy-Amsel-BV diagnosis and triangles indicate predicted NBV by proxy-Amsel-BV diagnosis for each sample. Each point is additionally colored by community state type. The plots are ordered left to right by linear model adjusted R^2 values between miRNA expression and Nugent score: 0.40 (miR-193b), 0.21 (miR-182), 0.18 (miR-203b), 0.14, (miR-378a), 0.13 (miR-3607), 0.12 (miR-324), 0.10 (miR-500a) and 0 (miR-146a).

Gene Ontology (GO) processes of experimentally validated miRNA targets for each of the common miRNAs (ranked by the number of GO processes per group) include transcription, cell growth and cell cycle, signaling, development, hypoxia and immune response (Figure 2.4, Table 2.2).

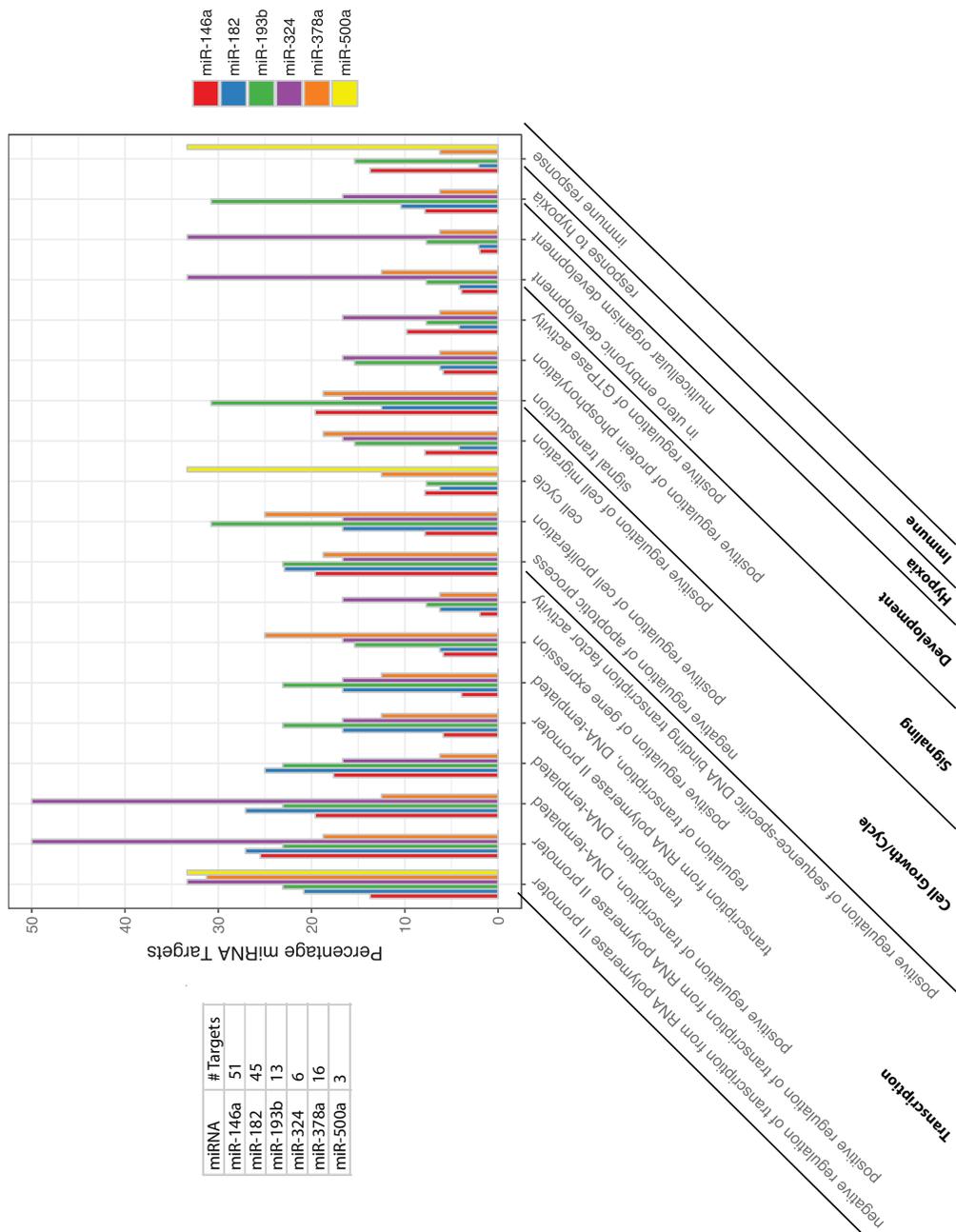


Figure 2.4 Gene Ontology processes of common significant miRNAs based on their experimentally validated gene targets.

Gene Ontology processes mapped to at least one miRNA gene target in at least four of the top miRNAs are grouped by involvement in transcription, cell cycle/growth, signaling, development, hypoxia or immunity. Table in upper left corner contains the number of gene targets per miRNA. miR-203b and miR-3607 did not have experimentally validated targets in miRTarBase.

Table 2.3 Gene Ontology processes for miR-193b based on the most number of experimentally validated targets.	
Gene Ontology Process	Validated Targets
GO:0007165~signal transduction	ESR1, NF1, PLAU, YWHAZ
GO:0001666~response to hypoxia	ETS1, NF1, PLAU, SMAD3
GO:0008284~positive regulation of cell proliferation	AKR1C2, ETS1, KRAS, SHMT2
GO:0006366~transcription from RNA polymerase II promoter	ESR1, ETS1, MAX
GO:0000122~negative regulation of transcription from RNA polymerase II promoter	CCND1, ESR1, SMAD3
GO:0006351~transcription, DNA-templated	CCND1, ESR1, SMAD3
GO:0042981~regulation of apoptotic process	ESR1, ETS1, MCL1
GO:0006355~regulation of transcription, DNA-templated	ESR1, MAX, SMAD3
GO:0045893~positive regulation of transcription, DNA-templated	ESR1, ETS1, SMAD3
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	ESR1, ETS1, SMAD3
GO:0001889~liver development	KRAS, NF1, SMAD3
GO:0032355~response to estradiol	CCND1, ESR1, ETS1
GO:0043066~negative regulation of apoptotic process	MCL1, SMAD3, YWHAZ

Overexpression of miR-193b is associated with NBV

miR-193b was one of the best annotated, highest ranked, and most highly expressed (\log_2 normalized reads >5 in all samples) miRNA in both Random Forest models. Its expression was negatively correlated with Nugent score and proxy-Amsel-BV diagnosis, and was therefore chosen for further study to determine its functional consequences (Figure 2.3, Table 2.3). The expression of miR-193b was validated using qPCR in a subset of samples representing NBV or PBV subjects, and normalized to a non-variable miRNA reference (hsa-RNU6). Using a mixed effect linear model, the ΔCt of miR-193b was increased by 1.4 ($p= 0.03$) in NBV samples relative to PBV, confirming the results found by small RNA-seq.

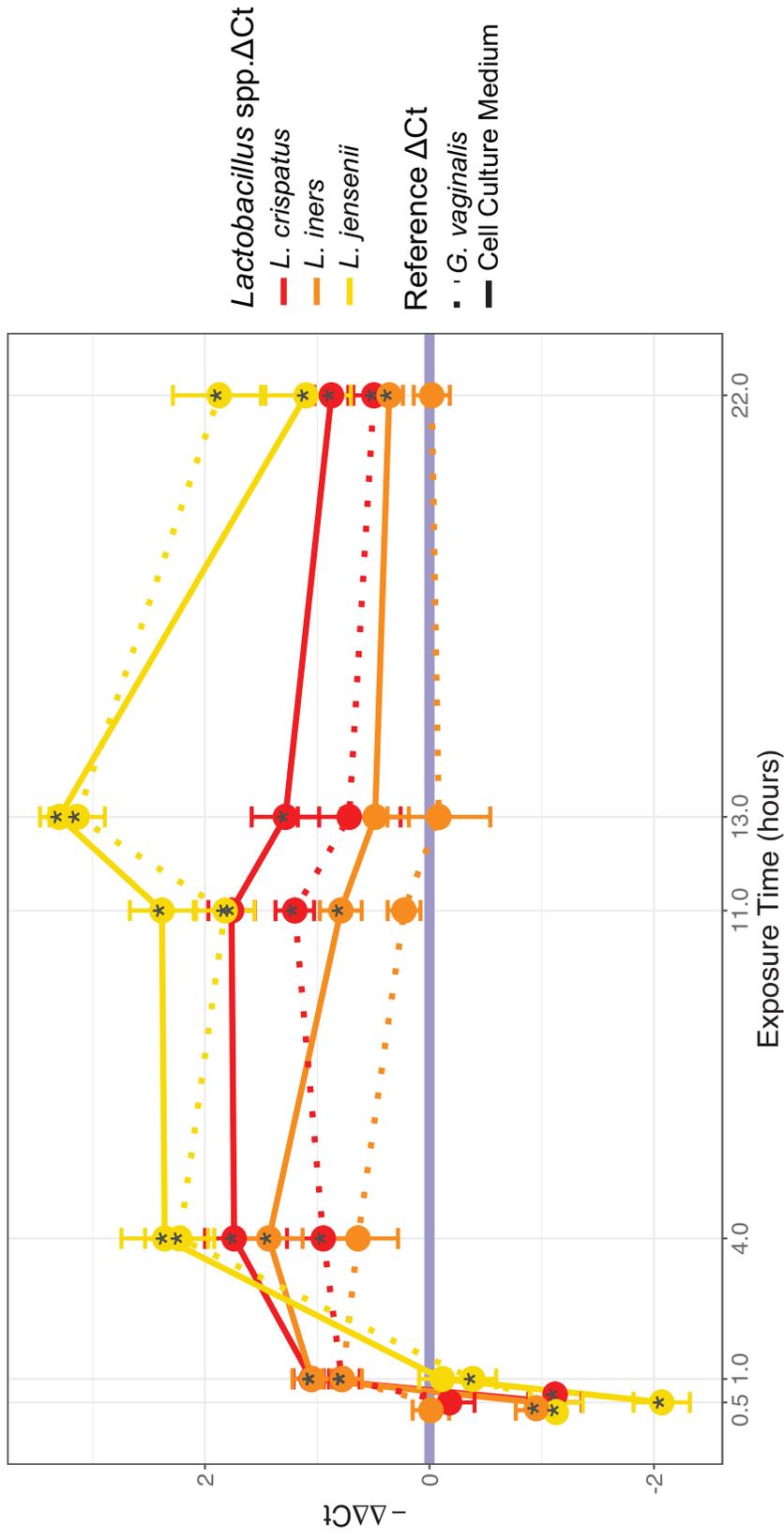


Figure 2.5 VK2 epithelial cell miR-193b expression time-course.

miR-193b expression after exposure to *L. crispatus*, *L. jensenii*, *L. iners*, *G. vaginalis* BCS or VK2 cell culture medium. miR-193b expression was quantified using ΔΔCt for *Lactobacillus* spp. relative to non-*Lactobacillus* (*G. vaginalis*) and cell culture medium sample references. Note that negative ΔΔCt values correspond to higher miR-193b relative expression of the *Lactobacillus* BCS exposed VK2 cell samples and so are shown as -ΔΔCt to match relationship of Ct and abundance. ΔΔCt corresponding to zero denoted with blue line. Statistically significant samples are noted with an “*” within the plot point ($p < 0.05$). Data are represented as mean \pm standard deviation.

Supernatant	Reference	Exposure Time	Delta delta Ct	pP-value	Significant
<i>L. crispatus</i>	<i>G. vaginalis</i>	0.5hr	0.178732	0.245022418	NA
<i>L. crispatus</i>	Cell Culture Medium	0.5hr	1.1185685	0.001421724	*
<i>L. jensenii</i>	<i>G. vaginalis</i>	0.5hr	1.128874333	0.002282757	*
<i>L. jensenii</i>	Cell Culture Medium	0.5hr	2.068710833	0.000209732	*
<i>L. iners</i>	<i>G. vaginalis</i>	0.5hr	0.012059333	0.904120521	NA
<i>L. iners</i>	Cell Culture Medium	0.5hr	0.951895833	0.001567842	*
<i>L. crispatus</i>	<i>G. vaginalis</i>	1hr	-0.781816	0.001082257	*
<i>L. crispatus</i>	Cell Culture Medium	1hr	-1.052936667	0.000335036	*
<i>L. jensenii</i>	<i>G. vaginalis</i>	1hr	0.386265	0.039104475	*
<i>L. jensenii</i>	Cell Culture Medium	1hr	0.115144333	0.397945898	NA
<i>L. iners</i>	<i>G. vaginalis</i>	1hr	-0.776163667	0.001601869	*
<i>L. iners</i>	Cell Culture Medium	1hr	-1.047284333	0.000511677	*
<i>L. crispatus</i>	<i>G. vaginalis</i>	4hr	-0.9462865	0.031002988	*
<i>L. crispatus</i>	Cell Culture Medium	4hr	-1.7403015	0.004530639	*
<i>L. jensenii</i>	<i>G. vaginalis</i>	4hr	-2.225599667	0.003519165	*
<i>L. jensenii</i>	Cell Culture Medium	4hr	-2.358896667	0.006296318	*
<i>L. iners</i>	<i>G. vaginalis</i>	4hr	-0.637242167	0.053215026	NA
<i>L. iners</i>	Cell Culture Medium	4hr	-1.431257167	0.002113527	*
<i>L. crispatus</i>	<i>G. vaginalis</i>	11hr	-1.200144	0.000427946	*
<i>L. crispatus</i>	Cell Culture Medium	11hr	-1.7636755	0.00050202	*
<i>L. jensenii</i>	<i>G. vaginalis</i>	11hr	-1.820595167	0.000698491	*
<i>L. jensenii</i>	Cell Culture Medium	11hr	-2.384126667	0.000171244	*
<i>L. iners</i>	<i>G. vaginalis</i>	11hr	-0.2266815	0.091781785	NA
<i>L. iners</i>	Cell Culture Medium	11hr	-0.790213	0.012239309	*
<i>L. crispatus</i>	<i>G. vaginalis</i>	13hr	-0.715701667	0.084691969	NA
<i>L. crispatus</i>	Cell Culture Medium	13hr	-1.283486333	0.002812047	*
<i>L. jensenii</i>	<i>G. vaginalis</i>	13hr	-3.138402	0.000211718	*
<i>L. jensenii</i>	Cell Culture Medium	13hr	-3.297535667	5.81E-06	*
<i>L. iners</i>	<i>G. vaginalis</i>	13hr	0.082877333	0.777175984	NA
<i>L. iners</i>	Cell Culture Medium	13hr	-0.484907333	0.056296014	NA
<i>L. crispatus</i>	<i>G. vaginalis</i>	22hr	-0.495473333	0.01026127	*
<i>L. crispatus</i>	Cell Culture Medium	22hr	-0.874167	0.000678162	*
<i>L. jensenii</i>	<i>G. vaginalis</i>	22hr	-1.874163	0.003754038	*
<i>L. jensenii</i>	Cell Culture Medium	22hr	-1.100347667	0.019514594	*
<i>L. iners</i>	<i>G. vaginalis</i>	22hr	0.020131333	0.84189057	NA
<i>L. iners</i>	Cell Culture Medium	22hr	-0.358562333	0.007923684	*
0.06% D-lactic acid	<i>G. vaginalis</i>	4hr	-0.079507333	0.65509605	NA
0.06% D-lactic acid	Cell Culture Medium	4hr	-0.223330667	0.178680764	NA
0.06% L-lactic acid	<i>G. vaginalis</i>	4hr	-0.122743	0.528962086	NA
0.06% L-lactic acid	Cell Culture Medium	4hr	-0.266566333	0.169632494	NA
0.06% L-lactic acid	0_06_D	4hr	-0.043235667	0.816123785	NA
<i>L. crispatus</i>	<i>G. vaginalis</i>	0.5hr	0.178732	0.245022418	NA
1% lactic acid, pH 7.66	<i>G. vaginalis</i>	4hr	-0.032712667	0.905280424	NA
1% lactic acid, pH 7.66	Cell Culture Medium	4hr	-0.022210333	0.936530483	NA
0.1% D-lactic acid	<i>G. vaginalis</i>	4hr	-1.530791	0.001908557	*
0.1% D-lactic acid	Cell Culture Medium	4hr	-1.520288667	0.002569712	*
0.1% D-lactic acid	0.1% L-lactic acid	4hr	-0.184439667	0.12271634	NA
0.1% L-lactic acid	<i>G. vaginalis</i>	4hr	-1.346351333	0.00627162	*
0.1% L-lactic acid	Cell Culture Medium	4hr	-1.335849	0.007659905	*

The relationship between miR-193b expression and the vaginal bacteria that typify various CSTs was further investigated *in vitro*. A monolayer of vaginal VK2 epithelial cells was grown in cell culture medium conditioned with 20% *L. crispatus*, *L. jensenii*, *L. iners* or *G. vaginalis* bacterial culture supernatant (BCS) for 0.5, 1, 4, 11, 13 or 22 hours, after which the relative expression ($\Delta\Delta\text{Ct}$) of miR-193b was quantified using qPCR (Figure 2.5, Table 2.4). The $\Delta\Delta\text{Ct}$ was computed by comparing one of the three *Lactobacillus* BCS to either VK2 cell culture medium alone or *G. vaginalis* BCS as a reference for each time point, i.e., $\Delta\text{Ct } L. \text{crispatus} - \Delta\text{Ct cell culture medium}$, $\Delta\text{Ct } L. \text{crispatus} - \Delta\text{Ct } G. \text{vaginalis}$, $\Delta\text{Ct } L. \text{jensenii} - \Delta\text{Ct cell culture medium}$, $\Delta\text{Ct } L. \text{jensenii} - \Delta\text{Ct } G. \text{vaginalis}$, $\Delta\text{Ct } L. \text{iners} - \Delta\text{Ct cell culture medium}$ and $\Delta\text{Ct } L. \text{iners} - \Delta\text{Ct } G. \text{vaginalis}$. The $-\Delta\Delta\text{Ct}$ of miR-193b in *L. jensenii* BCS versus cell culture medium exposed cells were highest overall, with the maximal difference of -3.3 at 13 hours post BCS exposure. Notably, all miR-193b $-\Delta\Delta\text{Ct}$ values except *L. iners* BCS relative to *G. vaginalis* BCS remain greater than 0 for the entire 22-hour time course not including the first hour. Interestingly, cells exposed to pH 7.66 buffered 1% DL lactic acid had inhibited cell proliferation similar to the *Lactobacillus* spp. conditions (7.2 +/- 1.8% filled scratch area, $p= 2.0 \times 10^{-2}$ when compared to *G. vaginalis* BCS and $p= 2.5 \times 10^{-2}$ when compared to cell culture medium), but miR-193b expression was not significant in these cells relative to non-*Lactobacillus* treatments (Figure 2.6, Table 2.4). Additionally, the $-\Delta\Delta\text{Ct}$ of miR-193b expression after 4 hours of exposure to 0.1% L-lactic acid relative to 0.1% D-lactic acid or 0.06% L-lactic acid relative to 0.06% D-lactic acid was not-significant (Figure 2.6, Table 2.4), but the expression of each isomer on miR-193b relative to

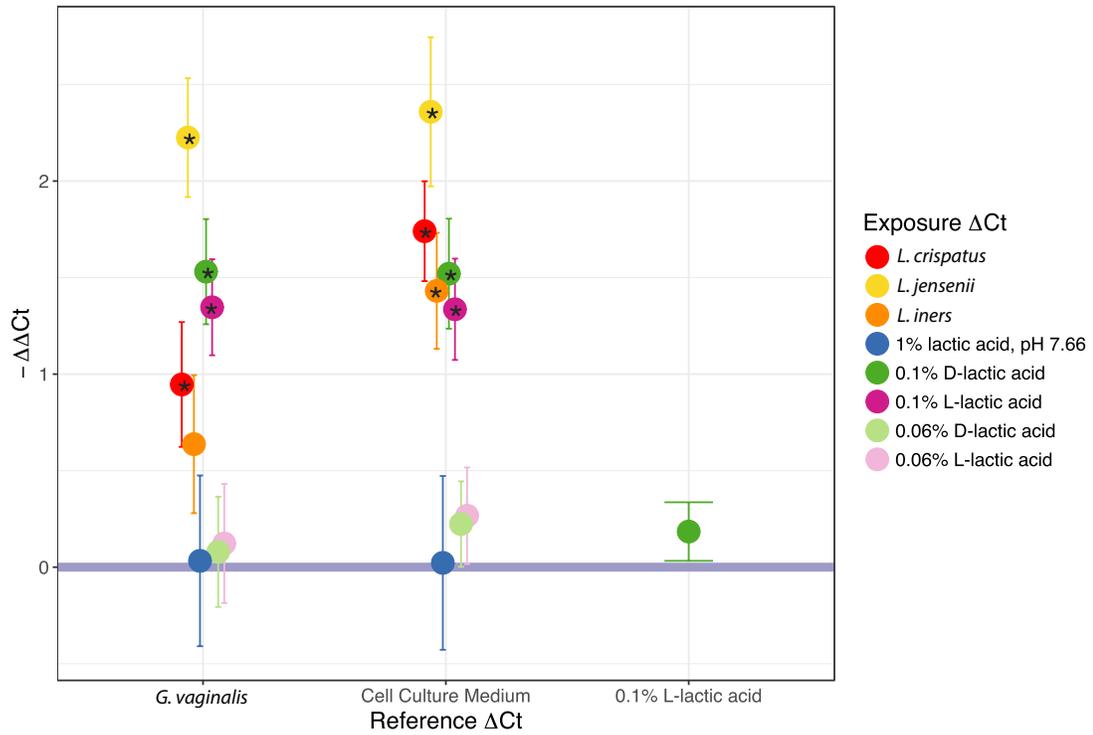


Figure 2.6 Relative expression of miR-193b following 4h exposure of D- or L-lactic acid.

Repeated are the 4h $-\Delta\Delta Ct$ expression value conditions Figure 2.5 in addition to D-lactic acid and L-lactic acid exposed VK2 cells. Statistically significant samples are noted with an “*” within the plot point ($p < 0.05$). Data are represented as mean \pm standard deviation.

either *G. vaginalis* BCS or cell culture medium reflected that of *Lactobacillus* spp. at 0.1% concentration (Figure 2.6, Table 2.4). The *in vitro* profiles reflect the *in vivo* results, where *Lactobacillus* spp. dominated vaginal microbiota are associated with higher miR-193b expression than those depleted of *Lactobacillus* spp. and typified by *G. vaginalis* BCS in this experiment.

Epithelial cell proliferation decreased when exposed to *Lactobacillus* spp. Bacterial Culture Supernatants

The function of miR-193b is broadly annotated as a tumor suppressor as miR-193b inhibits several genes associated with cell proliferation and metastasis in a variety of cell lines and mechanisms. A literature search identified the following transcripts as specific targets: tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta (YWHAZ), serine hydroxyl transferase (SHMT2), aldo-keto reductase family 1 member C2 (AKR1C2) [291], cell cyclin D1 (CCND1) [292], estrogen receptor- α (ESR1) [293], ETS proto-oncogene 1 transcription factor (ETS1) [294], KRAS proto-oncogene, GTPase (KRAS), MYC associated factor X (MAX) [295], neurofibromin 1 (NF1) [296], and urokinase-type plasminogen activator (PLAU/ PRAP1) [297]. Interestingly, miRNA-193b expression is also associated with increased cell proliferation in different cells or via different pathways (i.e. TGF- β signaling, targeting SMAD3 or PLAU in gastric cancer or human glioma, respectively) [298], [299]. The *in vivo* discovery of the relationship between the expression of miR-193b and different types of vaginal microbiota coupled with the *in vitro* reconstruction experiments validating these findings and the well-characterized functions of miR-193b, strongly suggested that miR-193b inhibits vaginal epithelial cell proliferation when exposed to metabolites produced by *Lactobacillus* spp. To test this directly, cell proliferation was quantified using a scratch assay and exposing VK2 cell monolayers to *L. crispatus*, *L. jensenii*, *L. iners*, or *G. vaginalis* BCSs for 13 hours, i.e. the time point at which the maximum effect of BCS on proliferation was empirically observed. All three *Lactobacillus* spp. BCS-

exposed VK2 cells displayed significant decreases in cell proliferation quantified by both proportion of cells present in the scratch area (Figure 2.7, Table 2.5) and cells synthesizing DNA (positive for 5-ethynyl-2'-deoxyuridine, EdU) (Figure 2.8, Table 2.5) relative to both *G. vaginalis* BCS and cell culture medium, coinciding with higher levels of miR-193b (Figure 2.5). There was also a significant decrease in filled scratch area and DNA synthesis in *G. vaginalis* BCS relative to VK2 culture medium (Figure 2.7, Figure 2.8, Table 2.5). Differences between *Lactobacillus* spp. BCS and *G. vaginalis* BCS were not explained by differences in cell proliferation between *Lactobacillus* NYC III culture medium alone and *G. vaginalis* culture medium alone (Figure 2.9). Taken together, the evidence suggests that a *Lactobacillus* spp.-dominated vaginal microbiota is associated with elevated miR-193b expression in vaginal epithelial cells relative to vaginal communities lacking *Lactobacillus* spp., resulting in decreased vaginal epithelial cell proliferation.

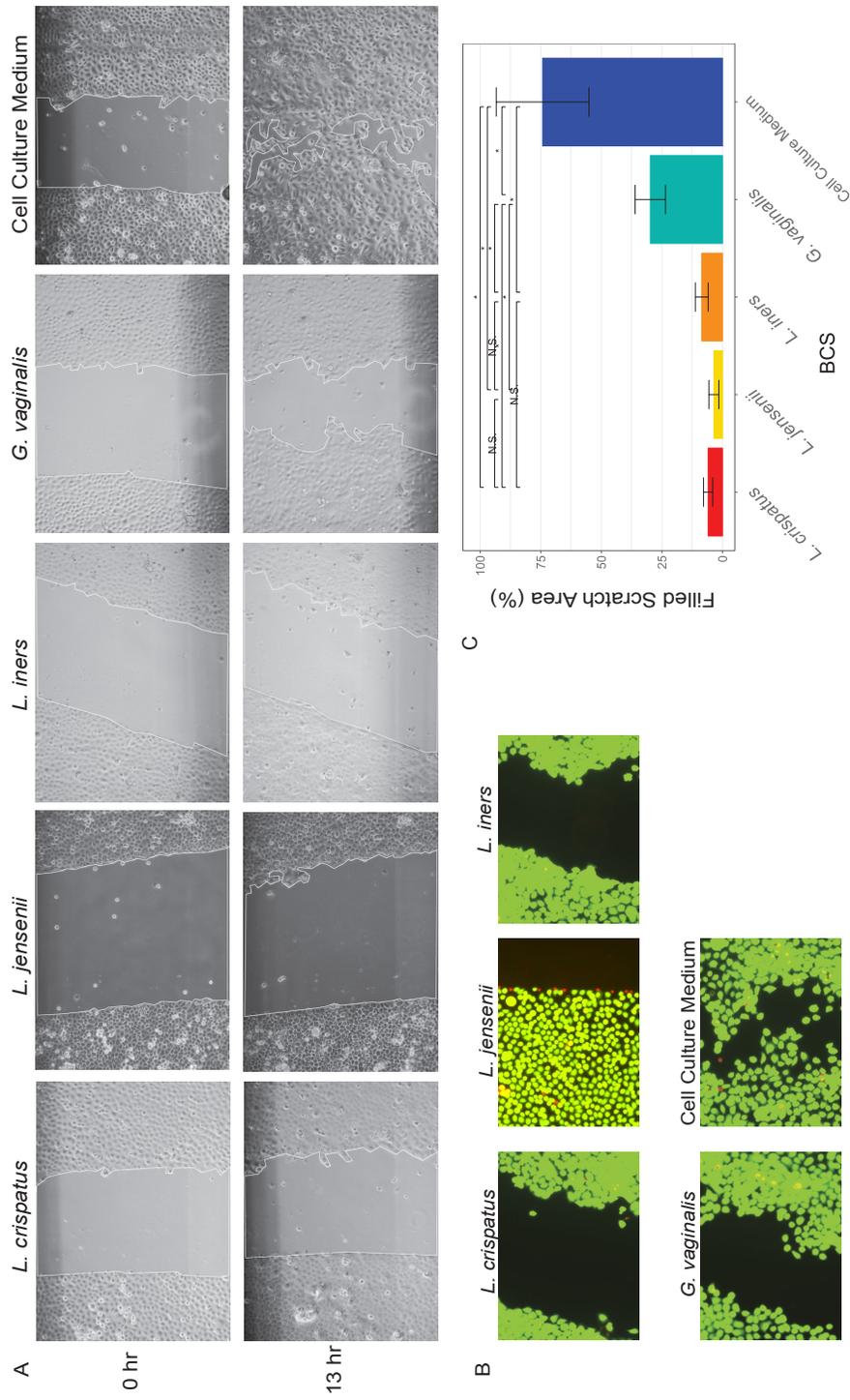


Figure 2.7 VK2 cell proliferation scratch assay, quantified by filled scratch area.

VK2 epithelial cells were scratched and exposed to respective BCS for 13h (A) Representative scratch assay microscopy at 100X of VK2 cells exposed to 20% *L. crispatus*, *L. jensenii*, *L. iners*, *G. vaginalis* BCS or cell culture medium at 0 and 13h post exposure; (B) VK2 cell viability following 13h BCS exposure (green is viable, red is non-viable). (C) Proportion of VK2 cells after 13h BCS or medium exposure filling the scratch area. Statistically significant comparisons ($p < 0.05$) are denoted with an “*” above the comparison line. Data are represented as mean \pm standard deviation.

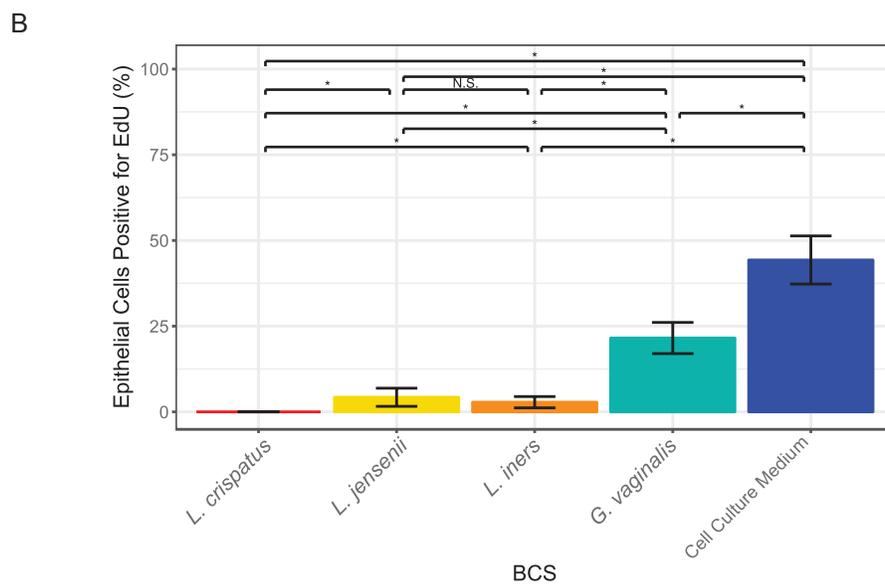
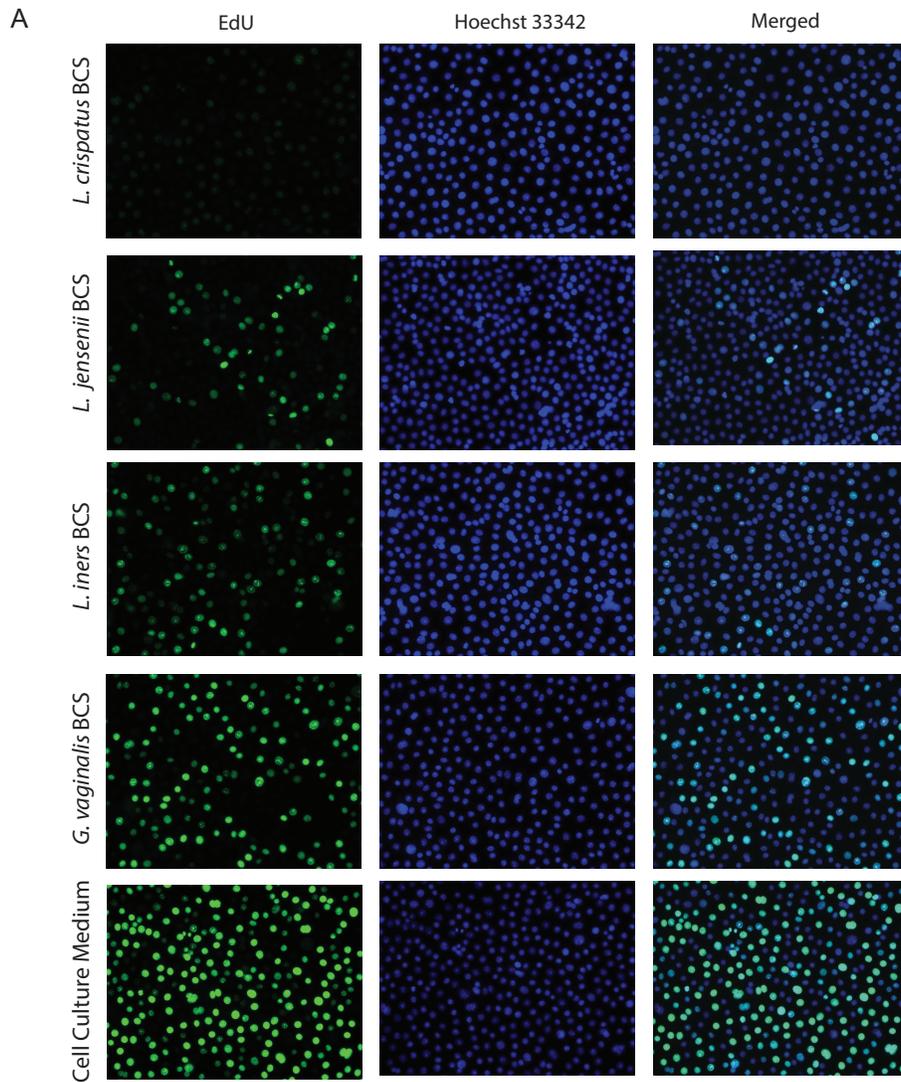
Table 2.5 Tabular data for cell proliferation assay				
<i>Lactobacillus</i> spp. BCS	Reference	p-value	Difference of means	Significant
<i>L. crispatus</i>	<i>L. jensenii</i>	0.213260924	-2.3829448	N.S.
<i>L. crispatus</i>	<i>L. iners</i>	0.241477195	2.616009733	N.S.
<i>L. crispatus</i>	<i>G. vaginalis</i>	0.015938865	23.86782947	*
<i>L. crispatus</i>	Cell Culture Medium	0.024237079	68.3016191	*
<i>L. jensenii</i>	<i>L. iners</i>	0.064219442	4.998954533	N.S.

Cell Cyclin D1 protein expression at 13 hours is reduced in vaginal epithelial cells exposed to *L. crispatus* and *L. jensenii* BCS but not *L. iners* BCS.

Cell Cyclin D1 (CCND1) protein expression in VK2 epithelial cells after 13-hour exposure to each BCS was assessed using Western blot immunofluorescence labeling (Figure 2.10). Relative to cell culture medium-exposed VK2 epithelial cells, the normalized expression of CCND1 protein was 85% and 75% less in *L. crispatus* and *L. jensenii* BCS-exposed VK2 cells, respectively, and 34% and 8% more in *L. iners* and *G. vaginalis* BCS-exposed VK2 cells, respectively. Thus, CCND1 production was decreased after exposure for 13h to *L. crispatus* and *L. jensenii* BCS, supporting that increased miR-193b expression is associated with reduced expression of at least CCND1.

Figure 2.8 VK2 cell proliferation scratch assay, quantified by EdU (following page)

(A) Representative fluorescence microscopy EdU Staining of Images taken at 100X for VK2s after 13-hour exposure to *L. crispatus* (top row), *L. jensenii* (second row) *L. iners* (third row), and *G. vaginalis* (fourth row) BCS and cell culture medium (fifth row). Left column shows EdU (green) staining for new DNA synthesis, middle column shows HOERST staining (blue) for nucleus and right column is the merged image. (B) Quantification of proportion of VK2 cells after 13 hours BCS or medium exposure positive for EdU nucleobases. Statistically significant comparisons ($p < 0.05$) are denoted with an “*” above the comparison line. Data are represented as mean +/- standard deviation.



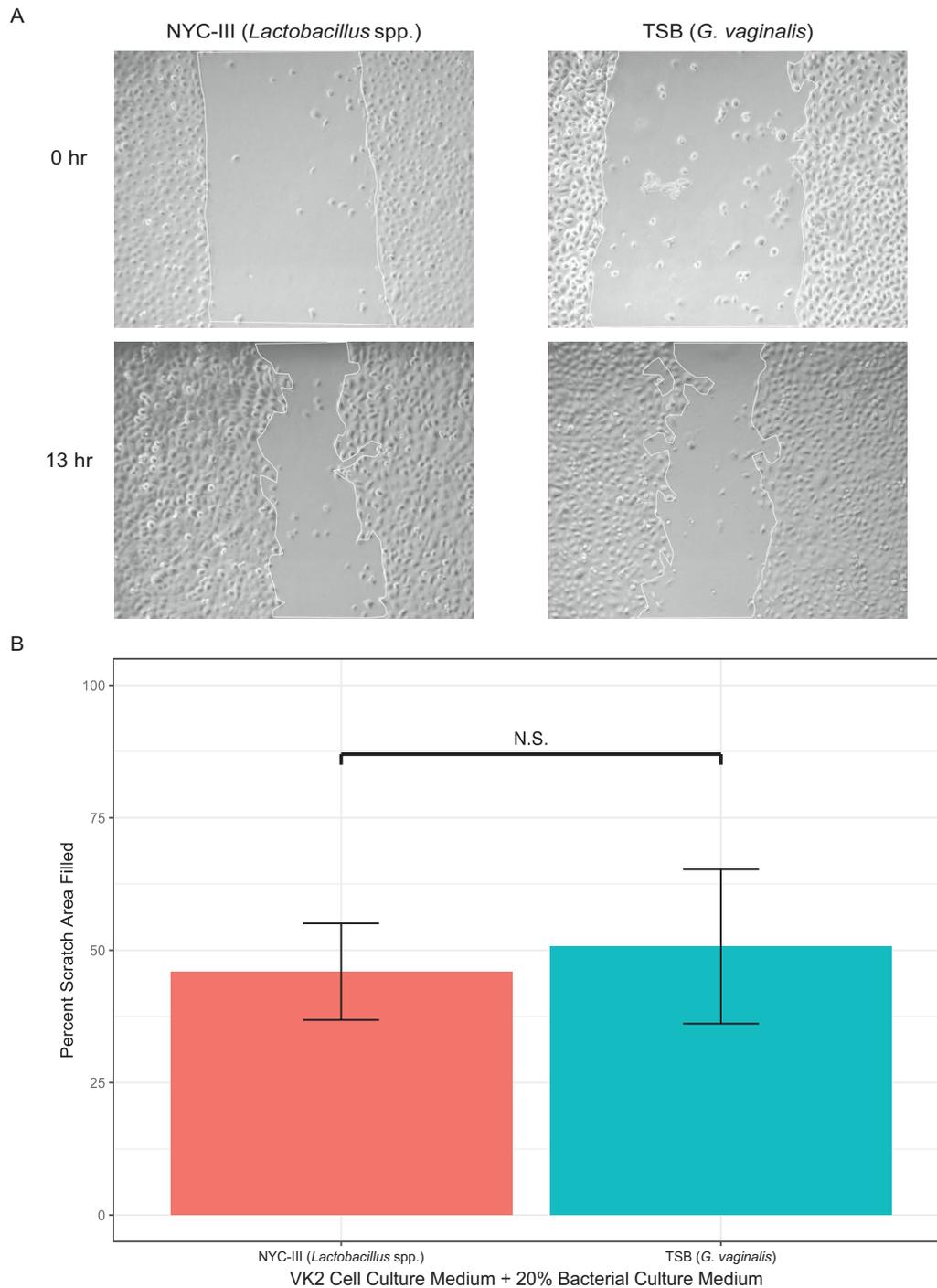


Figure 2.9 VK2 Cell proliferation scratch assay following 13-hour exposure to NYC-III (*Lactobacillus* spp.) or TSB (*G. vaginalis*) culture medium.

VK2 epithelial cells were scratched and exposed to respective bacterial culture medium for 13 hours. (A) Scratch assay microscopy at 100X of VK2 cells exposed to 20% NYC-III or TSB bacterial culture medium at 0 and 13 hours post exposure; (B) Proportion of VK2 cells after 13-hour medium exposure filling the scratch area. Proportion of filled scratch area is not statistically significant ($p=0.45$). Data are represented as mean \pm standard deviation.

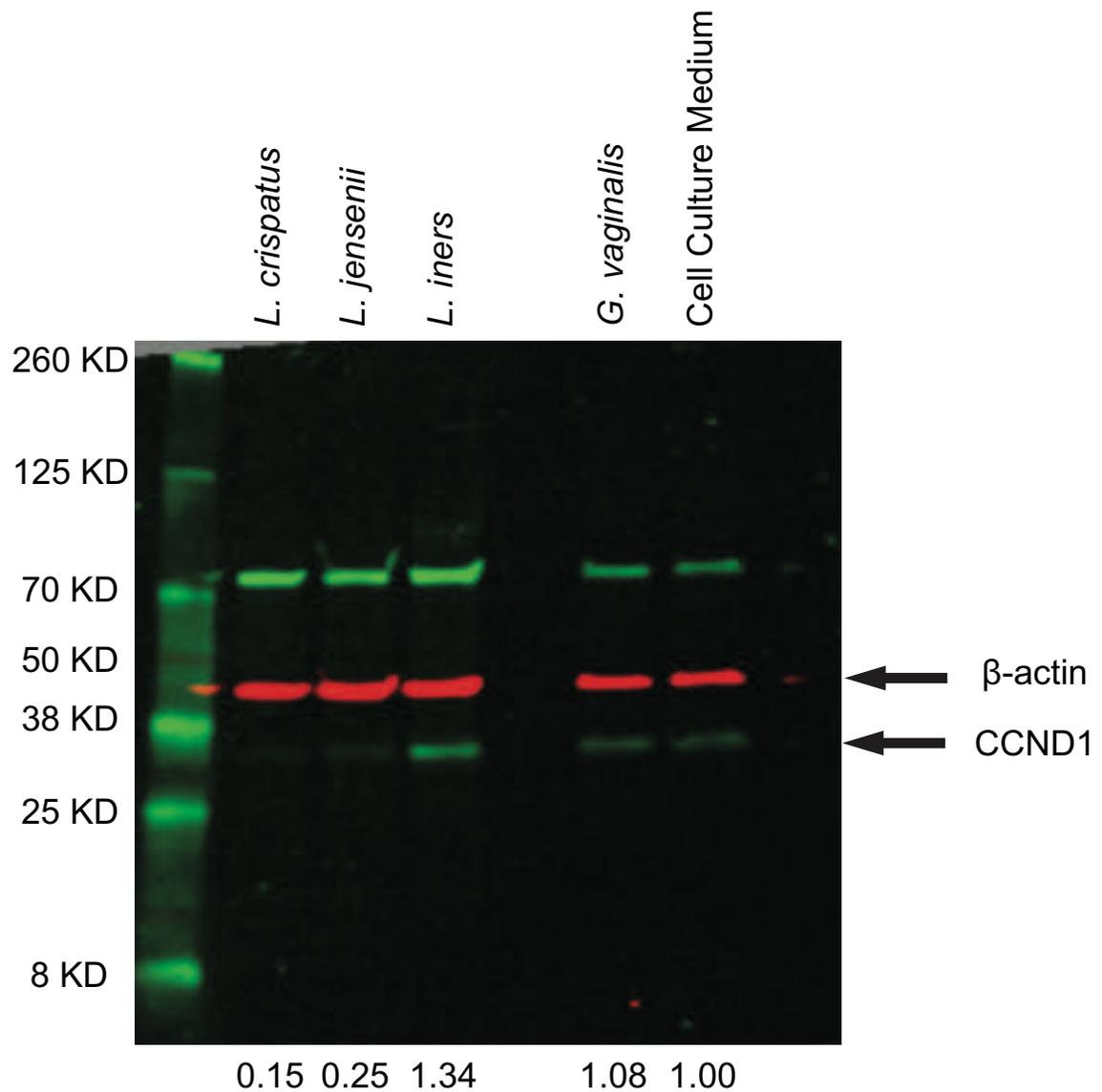


Figure 2.10 Western Blot of CCND1 after 13h exposure to BCS.

Protein expression of CCND1 (36 KDa) after 13h BCS exposure in VK2 cells (β -actin used as loading control, 42 KDa). Numbers under blot are normalized ratio of each BCS CCND1 intensity relative to VK2 cell culture medium. Quantification of cell images as percentage of cells. Band at ~70 KDa is an uncharacterized protein as indicated by the manufacturer (BD Pharmagen). Chameleon duo ladder (Li-Cor) loaded at a 1:1 ratio, 1 μ l/well

=

C. trachomatis infectivity is reduced in non-proliferating cervical epithelial cells

A2EN cervical epithelial cell proliferation was inhibited using one of two validated CDK4/cyclin D1 (proliferation) inhibitors, CAS 546102-60-7 [300] and Fascaplysin [301], and then challenged with *C. trachomatis*. Relative to culture medium-exposed cells, cervical epithelial cell proliferation evaluated by EdU incorporation, a marker of DNA synthesis, was decreased by 35.2% ($p=6.1 \times 10^{-5}$) and 28.5% ($p=3.5 \times 10^{-4}$) when cells were exposed to CAS 546102-60-7 and Fascaplysin, respectively (Figure 2.11, Table 2.6) After exposure of the epithelial cells to CAS 546102-60-7 and Fascaplysin, *C. trachomatis* infection was decreased by 53.4% ($p=3.3 \times 10^{-4}$) and 44.6% ($p=1.0 \times 10^{-3}$), respectively (Figure 2.11, Table 2.6). These results indicate that epithelial cell proliferation is required for efficient *C. trachomatis* infection (adjusted $R^2 = 0.94$, $p=2.6 \times 10^{-8}$).

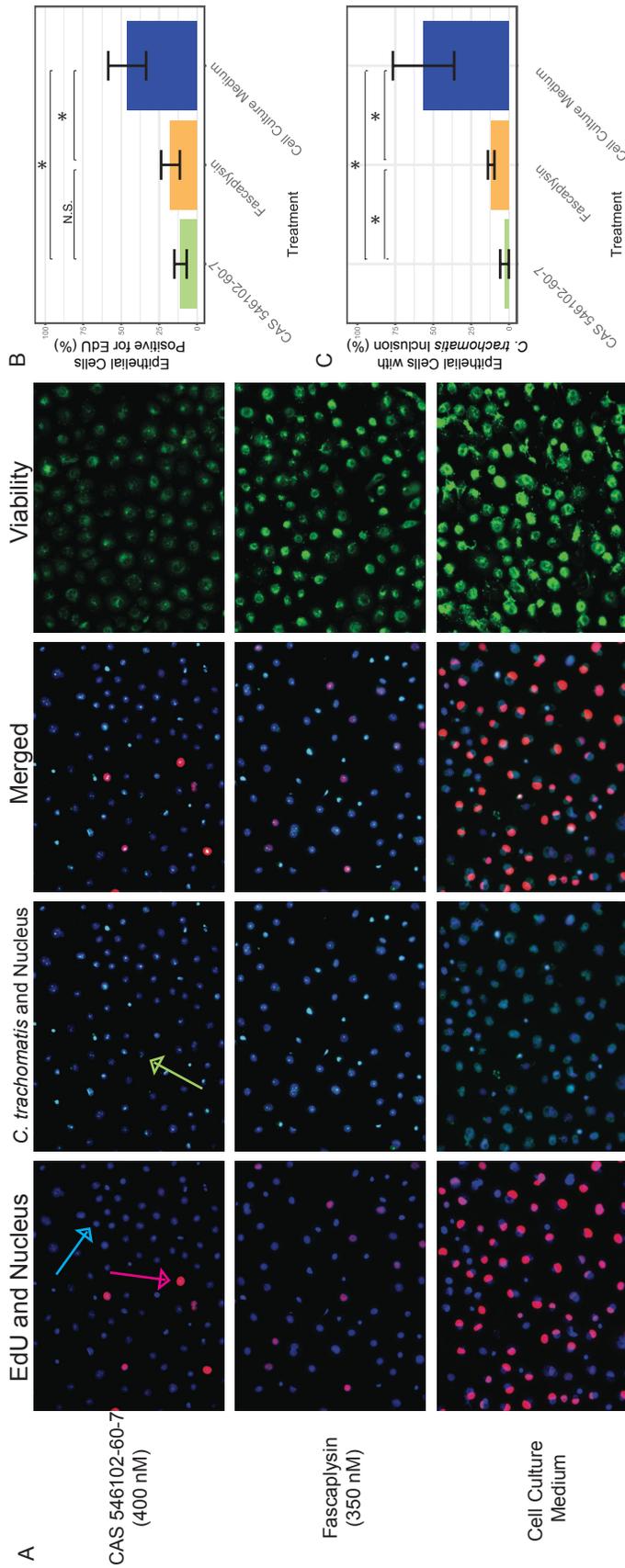


Figure 2.11 Effects of cell proliferation inhibition on *C. trachomatis* infectivity in A2EN human cervical cells.

(A) Representative fluorescence microscopy EdU and *C. trachomatis* infectivity staining following exposure to cell proliferation inhibitors. A2EN cervical epithelial cells exposed to CAS 546102-60-7 (top row), Fascaplysin (middle row) or cell culture medium (bottom row). Pink fluorescent staining is for EdU, green is for *C. trachomatis* and blue is HOERST staining for cell nuclei. Blue arrow shown in upper left panel is an example of non-proliferating cell, pink arrow shows an example of proliferating cell and green arrow in second panel shows *C. trachomatis*. Shown also is viability stain for each treatment (green is viable, red is non-viable). Quantification of cell images as percentage of cells (B) positive for EdU (proliferation) and (C) infected by *C. trachomatis* following exposure to inhibitor or medium. Statistically significant comparisons ($p < 0.05$) are denoted with an “*” above the comparison line. Data are represented as mean \pm standard deviation

Assay	Inhibitor	Reference	p-value	Mean difference	Significant
EdU	CAS 546102-60-7	Fascaplysin	0.149	-6.62	N.S.
EdU	CAS 546102-60-7	Cell Culture Medium	6.11E-05	-35.15	*
EdU	Fascaplysin	Cell Culture Medium	0.000353903	-28.53	*
<i>C. trachomatis</i> infectivity	CAS 546102-60-7	Fascaplysin	0.015253388	-8.80	*
<i>C. trachomatis</i> infectivity	CAS 546102-60-7	Cell Culture Medium	0.000334607	-53.39	*
<i>C. trachomatis</i> infectivity	Fascaplysin	Cell Culture Medium	0.001000266	-44.59	*

Discussion

This Aim characterized miRNA expression profiles of human vaginal samples associated with different vaginal microbiota community state types leveraging a large collection of specimens collected daily for 10 weeks and for which metataxonomic data and extensive metadata were available [19]. Among cellular functions known to be controlled by these miRNAs, epithelial cell proliferation was identified as a key mechanism impacted by changes in the vaginal microbiota with potential consequences on susceptibility and resistance to STI. *In vitro* experiments confirmed the *in vivo* findings by demonstrating the critical role of *Lactobacillus* spp. in maintaining the epithelial cells' protective homeostatic state. Lastly, the results show that chemically-induced arrest of epithelial cell proliferation is associated with reduced risk of *C. trachomatis* infection.

A total of 8 miRNAs had expression patterns that differed depending on the community state type of the vaginal microbiota. The ontology of genes targeted by these miRNAs included gene transcription, cell proliferation, migration, differentiation, apoptosis, development, immune response and response to hypoxia (Figure 2.4). miR-193b, whose expression was elevated concomitant with

Lactobacillus spp.-dominated vaginal microbiota and low Nugent scores (Figure 2.3) was selected for further analysis. miR-193b is known to control cell proliferation and has been implicated as a tumor suppressor by inhibiting cell proliferation and metastasis in a variety of cell lines (Table 2.3). This miRNA targets CCND1 [Table 2.3 [292]] and inhibits the G1 to S phase transition of the cell cycle. Production of CCND1 is indeed reduced *in vitro* when cells are exposed for 13h to culture supernatants of *L. crispatus* and *L. jensenii*, coincidental with a clear lack of cell proliferation in the scratch assay. Interestingly, a 13h exposure of VK2 epithelial cells to *L. iners* BCS resulted in increased in CCND1 expression, corresponding with an observed decrease in miR-193b expression from 11 to 13h (Figure 2.10, Figure 2.5), while maintaining a lack of proliferation (Figure 2.7, Figure 2.8). One would have expected increased proliferation, however two hours may have been insufficient for proliferation to be observed in the scratch assay resulting in the discordant relationship between increased CCND1 protein expression and the absence of proliferation. Additionally, *L. iners* may be unable to induce a sustained effect on proliferation, unlike *L. crispatus* and *L. jensenii*.

Bacterial culture supernatants were used in both the miR-193b qPCR assay and scratch assay experiments, and therefore strongly suggests that the observed effects are at least partially mediated by metabolites produced by *Lactobacillus* spp. Interestingly, two recent studies reported that gastrointestinal lactic acid bacteria and BV-associated bacteria both affect epithelial cell proliferation. Wound healing of HeLa cells was more significantly reduced upon 24-hour exposure to supernatants of *G. vaginalis* than by those of *L. iners* or the bacterial culture medium NYC-III [302].

In another study, lactate and acetate produced by *Lactobacillus casei* and *Bifidobacterium breve* inhibited gut epithelial cell proliferation that correlated with downregulation of cyclin D1 and cyclin E1 [211]. These studies provided mechanistic support for the interplay between microbiota and epithelial cell homeostasis and identify short chain fatty acids, such as lactate or acetate, as potential mediators. The results reported here are consistent with how *L. casei* and *B. breve* affect gastrointestinal cell proliferation, but interestingly inconsistent with *G. vaginalis* anti-proliferative effect on HeLa cells. It is likely that the cancerous cervical origin of the HeLa cell line makes it a poor model to study proliferation, since it probably carries mutations in cell cycle check point systems as suggested by [211]. The VK2 epithelial cell line used in our study (validated by genotyping; see analysis certificate in Appendix 3) is constructed similarly to the non-cancerous transformed murine intestinal crypt cell line m-ICc12 (eventually used by Matsuki et al. (2013)) and is therefore not expected to have large disruptions of cell cycle check points [303], [304]. Our findings are further supported by evidence that epithelial cell shedding is positively correlated with *G. vaginalis* and sialidase (produced by *G. vaginalis*) in a murine model and with high Nugent score [305]. This suggests that increased proliferation is a host defense mechanism against *G. vaginalis*, a BV-associated bacterial species known for its adherence and biofilm formation properties [306], [307]. In BV-associated states, increased epithelial cell shedding may aid in the host's ability to clear adherent bacterial cells from the epithelial surface giving *Lactobacillus* spp. the opportunity to colonize and/or adhere to the new epithelial surface, thus effectively displacing *G. vaginalis* from the vaginal epithelium (Figure

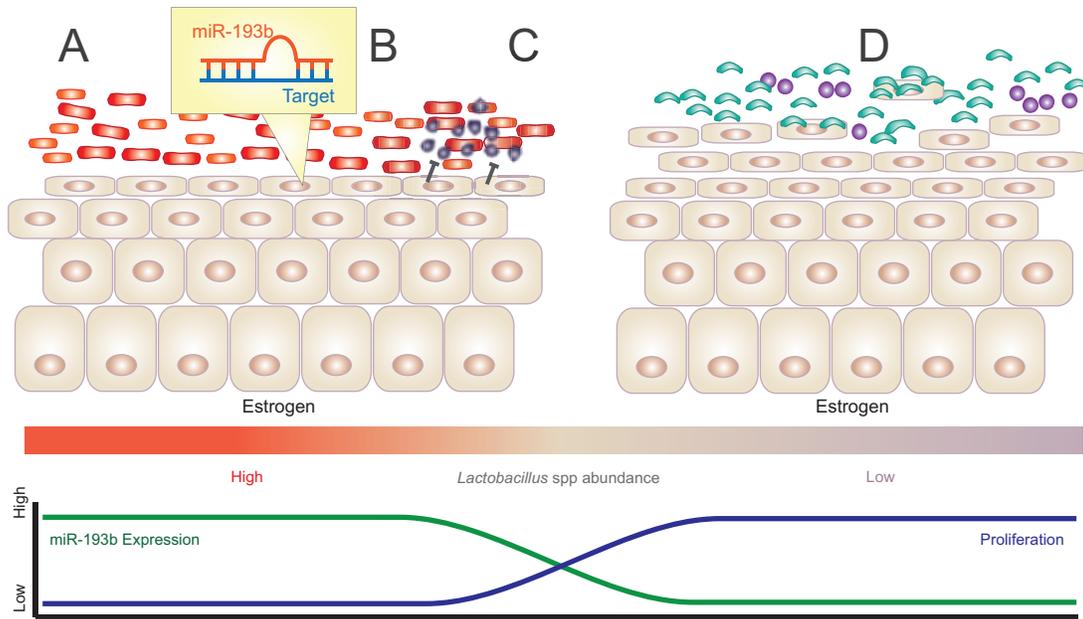


Figure 2.12 Vaginal epithelial cell homeostasis.

Characterization of vaginal epithelial cells revealed increased miR-193b expression and decreased cell proliferation when associated with *Lactobacillus* spp.-dominated relative to non-*Lactobacillus* spp. vaginal microbiota. Increased cell proliferation by estrogen and vaginal cell homeostasis may be influenced by (A) *Lactobacillus*-induced (red/orange shapes) miRNA-193b expression and targeting (B) leading to decreased cell proliferation and may result in a more stable microenvironment for *Lactobacillus* spp. to colonize by adhering to mucin or cell surface. Decreased cell proliferation protect from *C. trachomatis* (purple circles) infection (C), as demonstrated in the present study. Conversely, increased cell proliferation in the context of non-*Lactobacillus* spp.-dominated microbiota (oblong blue and circular purple shapes) result in cell shedding along with BV-associated microbes (D).

2.12) [308]-[310]. Other BV-associated bacterial species including *Peptoniphilus* spp., *Fusobacterium nucleatum* and *A. vaginae* have also been shown to adhere to ME-180 vaginal epithelial cells *in vitro* [307] and are predicted to induce a similar epithelial cell response. Consistent with these observations, these results demonstrated the differential impact of a vaginal microbiota dominated by *Lactobacillus* spp. or BV-associated species on vaginal epithelial cell proliferation.

The results additionally suggest *Lactobacillus* species-specific effects on epithelial cell responses as *L. jensenii* and *L. crispatus* exhibited the strongest positive effects on miR-193b expression, while *L. iners* was less pronounced, when compared to either *G. vaginalis* or cell culture medium (Figure 2.5). It was hypothesized

whether the differential production of L- and D-lactic acid in the three species could explain the observed differences. Whereas *L. crispatus* produces both the L and D isomers of lactic acid, *L. iners* and *L. jensenii* produce only the L and D forms, respectively [311]. The impact of lactic acid isomers on miR-193b expression was investigated by quantitative PCR (Figure 2.6). However, although both isomers caused increased miR-193b expression relative to *G. vaginalis* BCS or cell culture medium, VK2 epithelial cells were found to express miR-193b in equal abundance after 4 hours of exposure to D- or L-lactic acid. Thus, the small differences observed in miR-193b expression are not explained by differential production of lactic acid isomers and might be due to yet uncharacterized factors. Furthermore, while pH 7.66 buffered 1% DL lactic acid inhibited cell proliferation, miR-193b expression was not increased (Figure 2.6), suggesting that lactate and BCS may act through distinct mechanisms converging on similar pathways.

Estrogen has been shown to increase epithelial cell proliferation in the mouse and human vagina via growth factor signaling and the production of epidermal growth factor (EGF) receptor R1 (EGFR1) as well as estrogen receptor α (ER α), interestingly both of which are targets of miR-193b (Table 2.2) [312], [313]. The observed anti-proliferative effect of *Lactobacillus* spp. mediated by miR-193b may counterbalance the action of estrogen on cell proliferation, and ultimately contribute to the maintenance of vaginal epithelial cell homeostasis (Figure 2.12). *Lactobacillus* spp.-dominated microbiota controlled vaginal epithelial cell proliferation repression is speculated to be advantageous as it facilitates the rapid proliferative transition of the epithelium in response to BV-associated vaginal microbiota states *in vivo*.

Previous studies have shown that epithelial cell proliferation is activated upon infection by the sexually transmitted pathogen *C. trachomatis*, which requires EGFR for internalization and regulates cell proliferation by affecting cell transcription, DNA repair mechanisms, cyclin E, and PI3K, MEK and ERK growth signaling pathways [314]-[317]. The present study suggests that by controlling cell proliferation, the vaginal microbiota can counter the proliferation induced by the infecting pathogen and thus reduce susceptibility to infection. This is supported by the fact that BV-associated states are strong predictors of chlamydial infection among women who reported exposure to an infected partner. Conversely, women with a *Lactobacillus*-dominated vaginal microbiota are less likely to be infected [271]. The rate of transmission after contact with a partner infected by *C. trachomatis* is estimated to be 25-40% [318]-[320]. Given this, it was determined if decreased cell proliferation using chemical inhibitors of CDK4/cyclin D1 (key cell proliferation components) reduced *C. trachomatis* infectivity independent of the presence of *Lactobacillus* spp. Our results show that cervical epithelial cell proliferation is necessary for *C. trachomatis* to establish an infection. These results mechanistically point to cell proliferation controlled by the activity of the vaginal microbiota via modulation of miRNA expression as a major control center of vaginal epithelial homeostasis and thus protection against *C. trachomatis*. Further, because EGFR is a target of miR-146a and miR-21, two miRNAs identified in this study (Table 2.2), and is required for *C. trachomatis* EB internalization [316], it is likely that miRNA-driven regulation of EGFR contributes to the overall protection against *C. trachomatis* infection. We propose that a dynamic interplay between the vaginal microbiota, miRNA expression,

host cell proliferation and *C. trachomatis* infection exists as represented schematically on (Figure 2.12).

Limitations of the study. Random Forest models were used to discover miRNAs associated with *Lactobacillus* spp.-dominated states. The Nugent-BV Random Forest model on average predicted a correct Nugent score within 2 values while the proxy-Amsel-BV Random Forest model was 82.3% accurate predicting NBV and 80.3 % accurate predicting PBV. Although the accuracy is relatively high, Random Forest models can be difficult to interpret and generalize as they incorporate all available features for prediction. Thus, although the model provided acceptable accuracy and valuable insights in the discovery of BV-associated miRNAs and subsequently functional impact of the vaginal microbiota, it should not be used to predict outcomes of BV as this was not the goal of the analysis. The Nugent-RF and proxy-Amsel-RF predictive models were applied to increase the power to discover miRNA associated with Nugent-BV or proxy-Amsel-BV.

The expression of miR-193b was measured over a course of 22 hours in a VK2 epithelial monolayer cell model. Although this model was invaluable in its use to demonstrate concordant miR-193b expression between BCS and *in vivo* data, caution should be taken when generalizing to the *in vivo* vaginal environment. Namely, the effect of hormones (i.e. estrogen, progesterone), immune system (i.e. macrophages, leukocytes) and the three-dimensional matrix structure (i.e. basal vs. apical polarization) as well as microbial organisms (i.e. *Lactobacillus* spp.) are not fully accounted for and provide an additional layer of complexity to epithelial homeostasis in the vaginal environment.

Nonetheless, this aim provides clear evidence that *Lactobacillus*-dominated vaginal microbiota control the expression of miR-193b which in turns lowers CCND1 levels and reduce epithelial cell proliferation. More importantly, a low epithelial cell proliferative state confers resistance to *C. trachomatis* infection. Pathways associated with proliferative mechanisms should be further investigated as novel therapeutic targets to restore homeostasis of the vaginal microbiota and reduce the risk of *C. trachomatis* and other infections.

Chapter 3 *In vitro* vaginal epithelial cell transcriptional response to vaginal microbiota

Introduction

The cells that comprise the human vaginal epithelium sense and respond to *Lactobacillus* spp.-dominated and dysbiotic BV-associated microbiota (characterized by a paucity of *Lactobacillus* spp.) in part through the innate immune response [44]. A clear consensus on specific immune factors involved in the host response to dysbiotic vaginal microbiota is lacking, but some studies indicate that BV-associated microbes such as *Gardnerella vaginalis* induce IL-6 and IL-8, while *Lactobacillus crispatus* or *Lactobacillus jensenii* associated microbiota inhibit pro-inflammatory responses [41], [47], [59]. The beneficial effect of *Lactobacillus* spp. is provided through yet uncharacterized mechanisms. Chapter 2, however, strongly suggests that *Lactobacillus* spp. can decrease cell proliferation and thus maintain cell homeostasis. In some studies, *Lactobacillus iners* has been found to induce a slight inflammatory response, potentially due to the production of inerolysin, a pore-forming cholesterol-dependent cytolysin [42], [67], [100]. However, it is unclear if all strains of *L. iners* are equally capable of inducing an inflammatory response or express inerolysin *in vivo*. These observations raise questions regarding the ability for certain *Lactobacillus* species or strains to maintain optimal vaginal homeostasis and vaginal health.

Women harboring dysbiotic microbiota, as defined by Nugent score, have been shown to be at increased risk for the acquisition of STIs, including HIV, *Neisseria gonorrhoeae*, and *Chlamydia trachomatis* [28]-[30], [269]-[271] while women with a *Lactobacillus* spp.-dominated vaginal microbiota are less likely to be

infected [271]. The rate of transmission after contact with a partner infected by *C. trachomatis* is estimated to be 25-40% [318]-[320]. The current treatment for *C. trachomatis* infections includes the antibiotics azithromycin or doxycycline [321]. However, often the infection is asymptomatic and if left untreated, can ascend to the upper genital tract leading to long-term sequelae such ectopic pregnancy and infertility [128], [129]. Considering the significant burden this disease has on the healthcare system and women's health, novel protective measures are desperately needed. Building on major findings from Chapter 2, optimizing vaginal homeostasis will result in reduced risk to STI and can be achieved through research and translation of understanding the interaction between the host, the pathogen and the cervicovaginal microbiota. Such research could be leveraged to develop novel strategies to modulate and maintain a healthy vaginal microbiota.

In Chapter 2, both *in vivo* and *in vitro* longitudinal miRNA expression profiles of *Lactobacillus* spp.-exposed cells were used to elucidate miRNA-mediated gene regulation and function. Eight miRNAs were overexpressed in conditions where the vaginal microbiota was dominated by *Lactobacillus* spp. These miRNAs have experimentally validated targets associated with various gene ontology processes including transcriptional regulation, cell cycle, signaling, development, hypoxia, and immune response. When exposed to *G. vaginalis* Bacteria Culture Supernatant (BCS), a surrogate for dysbiotic vaginal microbiota, cultured vaginal epithelial cells were found to have decreased expression of miR-193b, which targets the G1-S phase cell-cycle checkpoint regulator, cyclin D1 (CCND1), resulting in increased cell proliferation. Increased cell proliferation was shown to significantly favor the ability

of *C. trachomatis* to infect cervical epithelial cells. Through the identification of miRNAs, a mechanism for the increased risk for *C. trachomatis* acquisition by dysbiotic microbiota has been established and suggests the host response to vaginal microbiota includes processes critical for susceptibility to *C. trachomatis* infection.

The physiological and biochemical processes associated with and affected by decreased vaginal epithelial cell proliferation resulting from *Lactobacillus* spp. exposure have not been characterized. An unaddressed question is how host cell sensing of the vaginal environment regulate miRNA expression. As this may be a critical defense mechanism unique to humans and their *Lactobacillus* spp.-dominated microbiota, there are likely multiple regulatory pathways that control the expression of miR-193b and other miRNAs.

This chapter characterizes the *in vitro* transcriptomic response of vaginal epithelial cells to *L. crispatus*, *L. jensenii*, *L. iners* and the BV-associated bacteria *G. vaginalis* BCS, as a surrogate for vaginal microbiota exposure over a time course sampled at 4h, 13h, and 22h, corresponding to increased miR-193b expression and decreased cell proliferation. Exposure to all three *Lactobacillus* spp. BCSs initially activated several immune-related pathways after 4h, however by 13h, *L. iners* and *G. vaginalis* were found to activate pro-inflammatory immune-related pathways while *L. crispatus* and *L. jensenii* BCSs minimally activated inflammatory pathways. In line with Chapter 2, *L. crispatus* and *L. jensenii* BCS decreased activation of cell cycle related pathways while *L. iners* did so moderately. Given the longitudinal differential expression patterns of key genes expressed by vaginal epithelial cells under these conditions, a model is proposed in which *Lactobacillus* spp., potentially through

metabolites such as lactate, inhibit histone deacetylases (HDACs) and other cell cycle regulator genes, leading to decreased cell cycle while the immunomodulation by *L. iners* and *G. vaginalis* activate immune pathways. The implications of this work suggest that vaginal epithelial cell proliferation is attenuated through global transcriptional changes mediated by exposure to the vaginal microbiota and open the possibility for novel therapeutics targeting vaginal epithelial cell proliferation to protect against STIs.

Methods

VK2 vaginal epithelial cell culture and bacterial culture supernatant (BCS) exposure

VK2 epithelial cells (ATCC CRL-2616, cell line authentication report in Appendix 3) were cultured at 37°C in 5% CO₂ then seeded at 7.5x10⁴ cells/well and grown to confluence in VK2 complete medium [Keratinocyte SFM, ThermoFisher # 17005042, with bovine pituitary extract (0.05 mg/ml), epidermal growth factor (0.1 ng/ml) and CaCl₂ (0.4 mM)]. BCSs were created by seeding 1x10⁷ bacteria/mL of either *L. crispatus* (ATCC 33197), *L. jensenii* (ATTC 25258), *L. iners* (ATTC 55195) or *G. vaginalis* (ATCC 14018) in 10 mL culture media (NYC-III for *L. crispatus*, *L. jensenii* and *L. iners*, TSB for *G. vaginalis*), grown anaerobically for 48 h, centrifuged at 3,000 x g for 10 minutes, sterile filtered (0.2µm filter) and stored at -20°C. BCSs were diluted to 20% (v/v) using complete VK2 cell culture medium and added to VK2 cells for a period of 4, 13 or 22h. VK2 cells were starved using base medium (Keratinocyte SFM only) for 24 h before adding VK2 cell culture medium containing 20% BCSs or VK2 cell culture medium alone. Immediately following the exposure

time, the cell culture medium was removed, and the cells were washed once with 1X PBS and 300 μ l RNAlater (QIAGEN) was added to wells. Cells were mechanically detached from plate and stored at -80°C for no more than 48 h before total RNA extraction.

Total RNA extraction from BCS exposed VK2 cells

Total RNA from VK2 vaginal epithelial cells exposed to BCS for 4, 13 and 22h or cell culture medium was extracted using the MasterPure™ Complete DNA and RNA Purification Kit (Epicentre, # MCR85102). BCS or cell culture medium exposed cells stored at -80°C in RNAlater were thawed and centrifuged for 10 minutes at 13,000 x g. The RNAlater was removed by pipetting, and cells were lysed with 300 μ l Lysis buffer containing 50 mg Proteinase K. The pellet was incubated for 15 minutes at 65°C with 10-second vortexing every 5 minutes, before placing the tubes on ice for 5 minutes. Following this step, 175 μ l Protein Precipitation buffer was added and the mixture vortexed for 10 seconds, and centrifuged for 10 minutes at 13,000 x g. The supernatant containing total nucleic acids was added to 500 μ l isopropanol and mixed by inverting the tube 40 times to precipitate nucleic acids, which were pelleted by centrifugation for 20 minutes at 13,000 x g, washed twice with 70% ethanol, and left to air-dry for 5-15 minutes. The pellet was resuspended in 10 μ l nuclease-free water (Ambion). DNA was then removed by adding 1 μ l TURBO DNA-free DNase (Life Technologies, # AM1907) and incubated for 30 minutes at 37°C followed by an additional treatment with 1 μ l of TURBO DNA-free DNase for another 30 minutes. DNase was inactivated by adding 2 μ l DNase Inactivation Buffer and incubated for 5 minutes at room temperature. The mixture was centrifuged

at 13,000 x g for 3 minutes and the supernatant (10 µl) containing total RNA transferred to a fresh tube. RNA quality and quantity were measured using 1 µl of RNA solution with a 2200 Agilent TapeStation and RNA screen tape (Agilent # 5067-5576). Samples were used to construct sequencing libraries the following day. All samples yielded 226 – 1,120 ng total RNA.

Ribosomal RNA-depleted (rRNA-depleted) RNA sequencing library construction

All rRNA-depleted RNA-seq libraries were prepared the same day to minimize batch effects and were carried out using the TruSeq Ribo-Zero Stranded Total RNA kit per manufacturer's recommendations (Illumina, # RS-122-2203) using 10 µl of total RNA as extracted above. For each sample, 5 µl of rRNA binding buffer and 5µl Ribo-Zero human/mouse/rat rRNA removal mix (Illumina) was added and then incubated at 68°C for 5 minutes. Following this step, the entire volume was added to 35 µl rRNA removal beads, incubated for 1 minute, and then beads were captured on a magnetic plate. The supernatant was mixed with 99 µl RNAClean XP beads (Beckman Coulter #A63987), incubated for 15 minutes, then placed on a magnetic plate where supernatant was removed and beads were washed once with 70 % ethanol. Beads were left to dry for 15 minutes, and 11 µl elution buffer was added. To elute the rRNA-depleted RNA, the beads were magnetized and the supernatant removed by pipetting. An 8.5 µl aliquot of the supernatant was added to 8.5 µl of the Elute, Prime, Fragment High mix (Illumina) and incubated at 94°C for 8 minutes, then placed on ice. First strand cDNA was performed by adding 8 µl of previously prepared reverse transcriptase mix (50 µl Superscript II (Life Technologies) into 450 µl First Strand Synthesis Actinomycete D (Illumina)) to the 17 µl of rRNA-depleted

RNA. The mixture was incubated at 25°C for 10 minutes, 42°C for 15 minutes, and then 70°C for 15 minutes before placing on ice. Second strand synthesis was carried out by adding 20 µl Second Strand Marking Master Mix (Illumina) to the mixture and incubating at 16°C for 1 hour. The reaction was cleaned using 90 µl of AMPure XP beads (Beckman Coulter #A63882), incubated for 15 minutes, then placed on magnetic plate and the supernatant discarded. The beads were cleaned using two washes with 80% ethanol, then air-dried for 15 minutes before adding 17.5 µl resuspension buffer. The beads were magnetized and the solution transferred to a fresh tube. To make the fragments compatible with adapters and prevent self-ligation, a 3'-adenosine overhang was added by adding 12.5 µl A-Tailing Mix (Illumina) to 15µl of the solution. The mixture was incubated at 37°C for 30 minutes, then 70°C for 5 minutes before being transferred to ice. Adapters were ligated by adding 2.5 µl Ligation Mix (Illumina) and 2.5 µl of a unique dual Illumina index to each sample. The mixture was incubated at 30°C for 10 minutes and then 5 µl Stop Ligation Buffer (Illumina) was added. The ligated cDNA was cleaned using AMPure XP bead clean-up by adding 42 µl of beads to the mixture, incubating for 15 minutes, placing mixture on magnetic stand, removing supernatant, and washing twice with 80% ethanol. The beads were resuspended in 52.5 µl Resuspension Buffer (Illumina), magnetized and 50 µl of the supernatant was added to 50 µl AMPure XP beads for 15 minutes. Beads were magnetized, the supernatant discarded, washed twice with 80% ethanol and dried for 15 minutes. The beads were resuspended in 22.5 µl Resuspension Buffer, magnetized, and 20 µl of the mixture transferred to a fresh tube for PCR enrichment. To selectively enrich DNA fragments that have adapter

molecules on both ends and to amplify the amount of DNA in the library 5 µl PCR Primer Cocktail (Illumina), 25 µl PCR Master Mix (Illumina) were added to the mixture and amplification was performed using 15 cycles at 98°C for 10 seconds, 60°C for 30 seconds and 72°C for 30 seconds, and a final extension at 72°C for 5 minutes. Enriched libraries were cleaned by adding 50 µl AMPure XP beads, incubating for 15 minutes, then placing the tubes on a magnetic stand and discarding the supernatant by pipetting. The beads were washed twice with 80% ethanol, air-dried for 15 minutes, and resuspended in 32.5 µl Resuspension Buffer. The beads were magnetized and 30 µl of the supernatant was transferred for subsequent library validation and sequencing. Libraries fragment size of 200-500bp range were validated on the LabChip GX (PerkinElmer).

RNA-seq library sequencing

RNA-seq libraries were sequenced on an Illumina HiSeq 4000 using the 150 bp paired-end protocol at the Institute for Genome Science's Genomic Resources Center (Baltimore, MD, USA). Indexed RNA-seq libraries were multiplexed at 15 samples per lane.

Sequence trimming, alignment and feature counting

RNA-seq reads were trimmed using trimmomatic version 0.33 using the following parameters: ILLUMINACLIP:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 (using Illumina adapter sequences, remove the first and last 3 bases below quality 3, with a 4 bp sliding window and trimming when quality drops below 15, and dropping reads below 36 bases long) [322]. Reads

were aligned to the hg19 human genome reference sequence using TopHat v2.1.0 with default parameters [323] (Human Genome Reference hg19 (available through Illumina at iGenomes ftp://illumina.com/Homo_sapiens/UCSC/hg19/Homo_sapiens_UCSC_hg19.tar.gz, downloaded August 17, 2015). Strand-specific genomic feature overlaps were counted using HTSeq version 0.5.3p3 [230] with default parameters (mode=union, minaaqual=0, stranded='yes') and the iGenomes annotation as above.

Read mapping quality control and differential expression

All analysis scripts can be found in Appendix 2. Sample replicates were validated by computing the \log_2 read count linear correlation coefficients between replicates. Samples with $R^2 < 0.9$ were excluded from further analysis, except where dropping a sample would result in a single sample per time point for a given treatment. To check for contamination, including the presence of human rRNA not aligned to the human reference, the top 10 most abundant unaligned reads per treatment were BLASTed against the non-redundant nucleotide collection to determine any non-human cross-contamination (from experimental sources or within-sequencing lane) or human rRNA contamination [324]) Samples having more than 90% human rRNA sequences were excluded from further analysis. All of the top 10 most abundant unaligned reads from all samples were of human origin.

The R package edgeR, version 3.10.5, was used to compute pairwise differential expression between combinations of each exposure time and BCS vs. cell culture medium [240]. Negative binomial dispersion was estimated for samples passing QC by applying the estimateDisp function available through edgeR, which

computes a common, trended, and gene-wise dispersion estimate to be used in downstream statistical inference. Samples were normalized using the `calcNormFactors` function which implements the Trimmed Mean of M-values (TMM) normalization procedure [242]. Reads were fit to a negative binomial generalized linear model using the `glmFit` function available in `edgeR`, using the sample's treatment as the design matrix. Differential expression using `edgeR`'s likelihood ratio test was computed for each gene using the `glmLRT` function. Genes with an average log counts per million ($\log\text{CPM}$) >1 , \log_2 -transformed Fold Change ($\log\text{FC}$) > 1 and false discovery rate (FDR) < 0.01 [325] were considered differentially expressed between treatments. The FDR corrects for multiple hypothesis testing [326]. The mean $\log\text{CPM}$ as calculated by `edgeR` is the \log_2 counts per million reads, averaged over all the libraries, while $\log_2\text{FC}$ is the coefficient of the Generalized Linear Model used by `edgeR` [240].

Pathway enrichment to identify commonly expressed pathways

Differentially expressed genes for each comparison were used to generate pathway enrichment scores using Ingenuity Pathway Analysis (IPA) (QIAGEN Inc., build version 439932M, content version 33559992). IPA computes a pathway activation score (z-score) for each pathway comparison based on inferred expression directionality using the $\log\text{FC}$ of each gene [327]. Pathway z-scores or $\log\text{FC}$ values were used to plot each comparison's time-course using the R library `ggplot2` (version 2.2.1) and custom scripts (see Appendix 2).

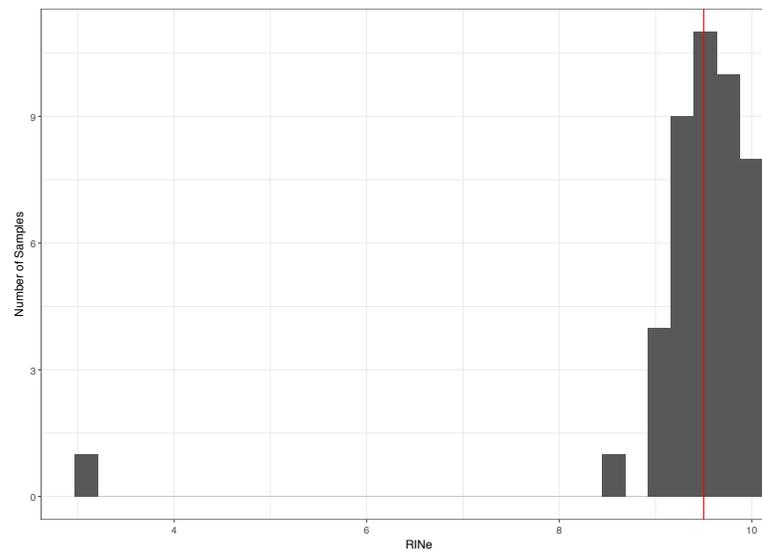


Figure 3.1 RINe distribution for RNA samples used in the study

Histogram of RINe quality scores for all extracted RNA samples used in study (45 samples). Red line indicates median RINe value 9.7. Single sample with a RINe of 3.1 failed QC.

Results

RNA-seq alignment statistics and quality control

Total RNA was extracted from triplicate VK2 cells exposed for 4h, 13h and 22h to either 20% (v/v) *L. crispatus*, *L. jensenii*, *L. iners*, or *G. vaginalis* BCS or cell culture medium. The median RINe quality score was 9.7 across 45 samples (range: 3.1-10) (Figure 3.1). Five of the initial 45 samples failed sequencing or library construction (*L. crispatus* BCS 4h exposure replicate 3, *L. jensenii* 4h exposure replicate 3, *L. iners* 4h exposure replicate 3, *G. vaginalis* 22h exposure replicate 1 and cell culture medium 22h exposure replicate 2). Sequencing and alignment statistics are shown in Table 3.1 and raw read counts are in Appendix 11. *L. crispatus* BCS 13h exposure replicate 3 sample had a relatively high proportion (94.3%) of human rRNA

reads and cell culture medium 4h exposure replicate 3 sample was poorly correlated ($R^2 < 0.9$) with replicate 1 and 2 samples. Both samples were therefore removed, leaving 38 samples for subsequent analysis. The top 10 most abundant unaligned reads from each of the remaining sample were found to be human rRNA that was not efficiently removed during library construction.

Vaginal epithelial immune response and cell cycle pathways are associated with BCS exposure

Differentially expressed genes were identified for each exposure time point using each BCS treatment (*L. crispatus*, *L. jensenii*, *L. iners* or *G. vaginalis*) compared to cell culture medium (Appendix 12). The number of differentially expressed genes, computed by edgeR and defined as genes having False Discovery Rate (FDR) corrected p-value < 0.01 , mean \log_2 (Counts Per Million (CPM)) > 1 and \log_2 fold change (FC) > 1 , for each comparison group are shown in Table 3.2.

For each BCS vs. cell culture medium comparison, Ingenuity Pathway Analysis (IPA) software computed canonical pathway activation scores (z-scores) based on the expression of differentially expressed genes (Appendix 13). The z-score is a way of assessing the agreement between each pathway's predicted and observed gene activation relationships, where positive z-scores correspond to pathways predicted to be activated given the direction (positive/negative) \log_2 -transformed Fold Change (logFC) expression values and known gene-gene regulatory interactions within the pathway (and conversely, negative z-scores correspond to repressed pathway activation) [327]. Among the most 28 most common activated or repressed canonical pathways (defined as having absolute z-score greater than 2 in at least

Table 3.1 Read alignment statistics for VK2 cells exposed to each BCS treatment for 4, 13, and 22h

Treatment	Exposure Time (hours)	Replicate	Number of reads after trimming	Number of reads aligned to hg19	Percent aligned to hg19
<i>L. crispatus</i> BCS	4	1	14,867,552	5,558,855	37.4
<i>L. crispatus</i> BCS	4	2	6,325,764	1,648,935	26.1
<i>L. crispatus</i> BCS	13	1	7,782,296	3,386,120	43.5
<i>L. crispatus</i> BCS	13	2	7,752,780	3,024,389	39.0
<i>L. crispatus</i> BCS	13	3	2,829,593	161,415	5.7*
<i>L. crispatus</i> BCS	22	1	5,510,460	1,416,161	25.7
<i>L. crispatus</i> BCS	22	2	11,631,321	4,392,899	37.8
<i>L. crispatus</i> BCS	22	3	21,444,118	8,448,815	39.4
<i>L. jensenii</i> BCS	4	1	6,891,476	1,533,479	22.3
<i>L. jensenii</i> BCS	4	2	7,347,943	3,206,459	43.6
<i>L. jensenii</i> BCS	13	1	6,298,882	1,921,098	30.5
<i>L. jensenii</i> BCS	13	2	8,055,876	2,974,822	36.9
<i>L. jensenii</i> BCS	13	3	10,203,941	3,802,258	37.3
<i>L. jensenii</i> BCS	22	1	7,477,077	2,950,478	39.5
<i>L. jensenii</i> BCS	22	2	7,388,378	2,049,302	27.7
<i>L. jensenii</i> BCS	22	3	20,070,883	8,633,553	43.0
<i>L. iners</i> BCS	4	1	7,857,237	2,645,213	33.7
<i>L. iners</i> BCS	4	2	7,907,060	3,115,872	39.4
<i>L. iners</i> BCS	13	1	4,323,103	863,055	20.0
<i>L. iners</i> BCS	13	2	4,547,722	1,729,117	38.0
<i>L. iners</i> BCS	13	3	14,951,198	6,838,588	45.7
<i>L. iners</i> BCS	22	1	4,303,383	1,532,055	35.6
<i>L. iners</i> BCS	22	2	4,508,819	1,021,441	22.7
<i>L. iners</i> BCS	22	3	12,587,497	5,445,716	43.3
<i>G. vaginalis</i> BCS	4	1	8,861,783	3,555,610	40.1
<i>G. vaginalis</i> BCS	4	2	4,989,044	1,880,874	37.7
<i>G. vaginalis</i> BCS	4	3	12,819,802	5,603,541	43.7
<i>G. vaginalis</i> BCS	13	1	8,291,229	3,483,930	42.0
<i>G. vaginalis</i> BCS	13	2	8,136,856	3,155,228	38.8
<i>G. vaginalis</i> BCS	13	3	22,489,737	10,538,780	46.9
<i>G. vaginalis</i> BCS	22	2	4,993,600	1,354,926	27.1
<i>G. vaginalis</i> BCS	22	3	13,341,795	5,545,096	41.6
Cell culture medium	4	1	9,492,111	3,745,190	39.5
Cell culture medium	4	2	8,385,668	3,207,936	38.3
Cell culture medium	4	3	10,862,411	5,132,964	47.3*
Cell culture medium	13	1	6,133,669	2,030,629	33.1
Cell culture medium	13	2	11,165,638	4,892,080	43.8
Cell culture medium	13	3	13,234,638	4,873,897	36.8
Cell culture medium	22	1	11,354,674	4,670,961	41.1
Cell culture medium	22	3	14,132,637	7,577,938	53.6

*Denotes sample was dropped due to QC failure

Table 3.2 Number of differentially expressed genes per pairwise comparison		
BCS comparison vs. cell culture medium	Exposure Time (hours)	Number of DE Genes
<i>L. crispatus</i> BCS vs. cell culture medium	4	1022
<i>L. jensenii</i> BCS vs. cell culture medium	4	868
<i>L. iners</i> BCS vs. cell culture medium	4	838
<i>G. vaginalis</i> BCS vs. cell culture medium	4	246
<i>L. crispatus</i> BCS vs. cell culture medium	13	4098
<i>L. jensenii</i> BCS vs. cell culture medium	13	2753
<i>L. iners</i> BCS vs. cell culture medium	13	436
<i>G. vaginalis</i> BCS vs. cell culture medium	13	122
<i>L. crispatus</i> BCS vs. cell culture medium	22	5505
<i>L. jensenii</i> BCS vs. cell culture medium	22	3039
<i>L. iners</i> BCS vs. cell culture medium	22	585
<i>G. vaginalis</i> BCS vs. cell culture medium	22	166

one comparison) were pathways related to the cell cycle (11 pathways, annotated by IPA) or immunity/proinflammation (12/7 pathways) (Figure 3.2 and Figure 3.3). The remaining 5 pathways did not belong to immunity nor cell cycle (LXR/RXR Activation, NRF2-mediated Oxidative Stress Response, PPAR Signaling, Type I Diabetes Mellitus Signaling and UVA-Induced MAPK Signaling).

When compared to cell culture medium exposed cells after 4h, 13h or 22h exposure, *L. iners* and *G. vaginalis* BCS generally activated immune pathways in VK2 cells while *L. crispatus* BCS generally activated immune pathways only at the 4h exposure (Figure 3.2). Specifically, comparisons with 4h BCS exposure relative to cell culture medium had the greatest number of immune-related pathways with z-score > 2: *G. vaginalis* vs. cell culture medium (18 pathways, 8 of which are pro-inflammatory), *L. iners* vs. cell culture medium (16 pathways, 7 of which are pro-inflammatory), and *L. crispatus* vs. cell culture medium (11 pathways, 5 of which are

Table 3.3 Number of immune pathways (Figure 3.2) and cell cycle pathways (Figure 3.3) with absolute z-score greater than 2. Number of pro-inflammatory pathways within immune pathways in parenthesis.					
BCS comparison vs. cell culture medium	Exposure Time (hours)	Cell cycle pathways		Immune pathways (pro-inflammatory)	
		Number z>2	Number z<-2	Number z>2	Number z<-2
<i>L. crispatus</i> BCS vs. cell culture medium	4	3	2	11 (5)	0 (0)
<i>L. jensenii</i> BCS vs. cell culture medium	4	1	1	2 (2)	0 (0)
<i>L. iners</i> BCS vs. cell culture medium	4	4	0	16 (7)	0 (0)
<i>G. vaginalis</i> BCS vs cell culture medium	4	3	1	18 (8)	0 (0)
<i>L. crispatus</i> BCS vs. cell culture medium	13	1	4	2 (1)	0 (0)
<i>L. jensenii</i> BCS vs. cell culture medium	13	0	2	1 (1)	1 (0)
<i>L. iners</i> BCS vs. cell culture medium	13	0	0	7 (4)	0 (0)
<i>G. vaginalis</i> BCS vs cell culture medium	13	1	0	8 (3)	0 (0)
<i>L. crispatus</i> BCS vs. cell culture medium	22	0	3	0 (0)	1 (0)
<i>L. jensenii</i> BCS vs. cell culture medium	22	0	2	1 (1)	1 (0)
<i>L. iners</i> BCS vs. cell culture medium	22	0	1	3 (1)	0 (0)
<i>G. vaginalis</i> BCS vs cell culture medium	22	0	0	2 (1)	0 (0)

pro-inflammatory), while *L. jensenii* vs. cell culture medium at the 4h exposure had only 2 immune-related and 2 pro-inflammatory pathways with z-score > 2 (Table 3.3, Appendix 13). Conversely, relative to cell culture medium, *L. crispatus* and *L. jensenii* BCS generally repress cell proliferation pathways at 4h, 13h or 22h exposure (Figure 3.3, Table 3.3, Appendix 13). These patterns demonstrate that the VK2 cell host responses to *L. crispatus* BCS and *L. jensenii* BCS are different from that of *L. iners* BCS or *G. vaginalis* BCS, which instead trigger similar immune related pathways while ineffectively repressing cell proliferation pathways.

An observation from Figure 3.2, Figure 3.3 and Table 3.3 is that *L. crispatus* BCS comparisons tend to both activate immune related pathways and suppress cell cycle related pathways while the same contrasting pattern is not as apparent for *L. jensenii* or *L. iners* BCS comparisons. At 4h after exposure, *L. crispatus* BCS increased 11 immune-associated pathways relative to cell culture medium, and of those, 5 are classified as pro-inflammatory pathways. At the same time, *L. crispatus* BCS exposed cells were among comparisons with the greatest number of decreased

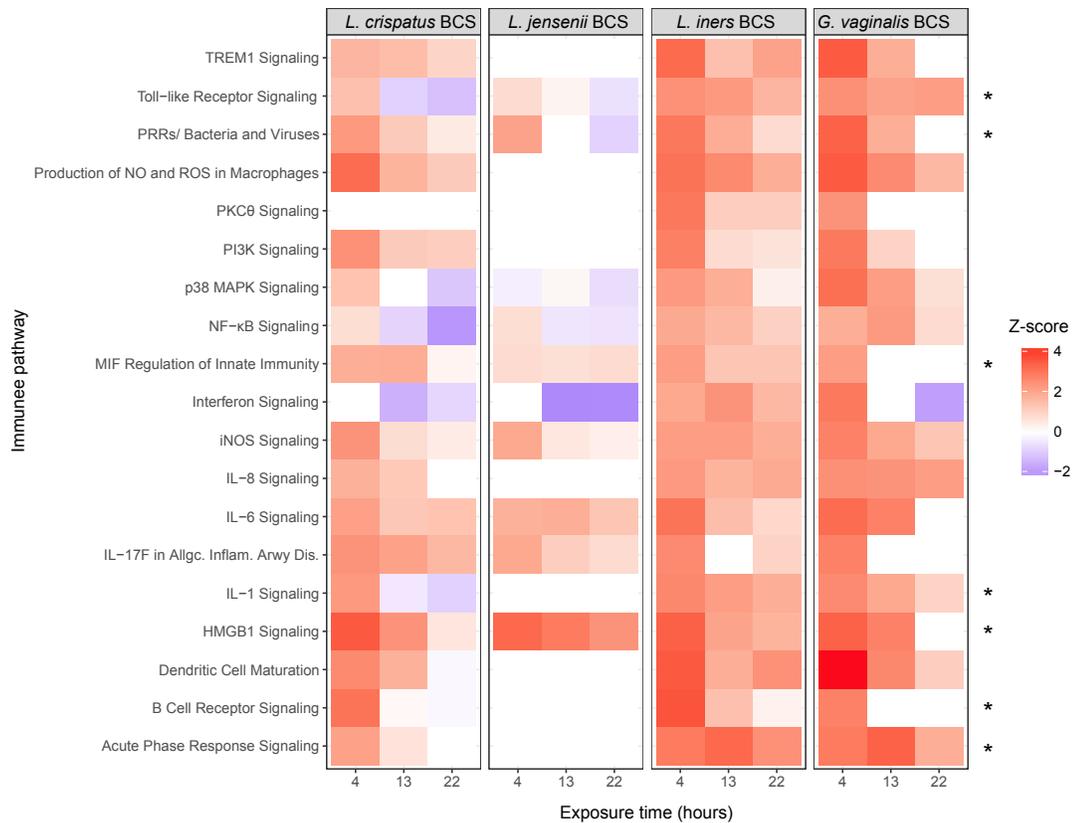


Figure 3.2 Heat map of activation scores (z-score) from pathways associated with immunity for cells exposed to each BCS for 4h, 13h or 22h vs. cell culture medium.

Pathways with absolute z-score greater than 2 in at least 1 comparison revealed decreased activation of immune response pathways in *L. crispatus* and *L. jensenii* BCS exposed cells. Conversely, *L. iners* and *G. vaginalis* BCS exposed cells maintain relatively high activation of immune pathways at 13h and 22h exposure. Pathways denoted with “*” are proinflammatory response pathways.

cell cycle expression pathways in 13h and 22h BCS exposure (Table 3.3). The longitudinal logFC expression patterns of IL6 and IL8 (Figure 3.4), cytokines that have been previously shown to be involved in the immune response to vaginal bacteria, both increase over time in *L. crispatus* comparisons while the remaining *Lactobacillus* spp. comparisons do not show as large of a shift in IL-8 logFC expression and show a decreased IL-6 logFC expression after 13h of BCS exposure. However, *L. crispatus* BCS only moderately activates the IL-6 and IL-8 signaling pathways (Figure 3.2), with only one comparison having a z-score > 2 (IL-6 signaling pathway at 4h after *L. crispatus* BCS exposure relative to cell culture medium. Figure

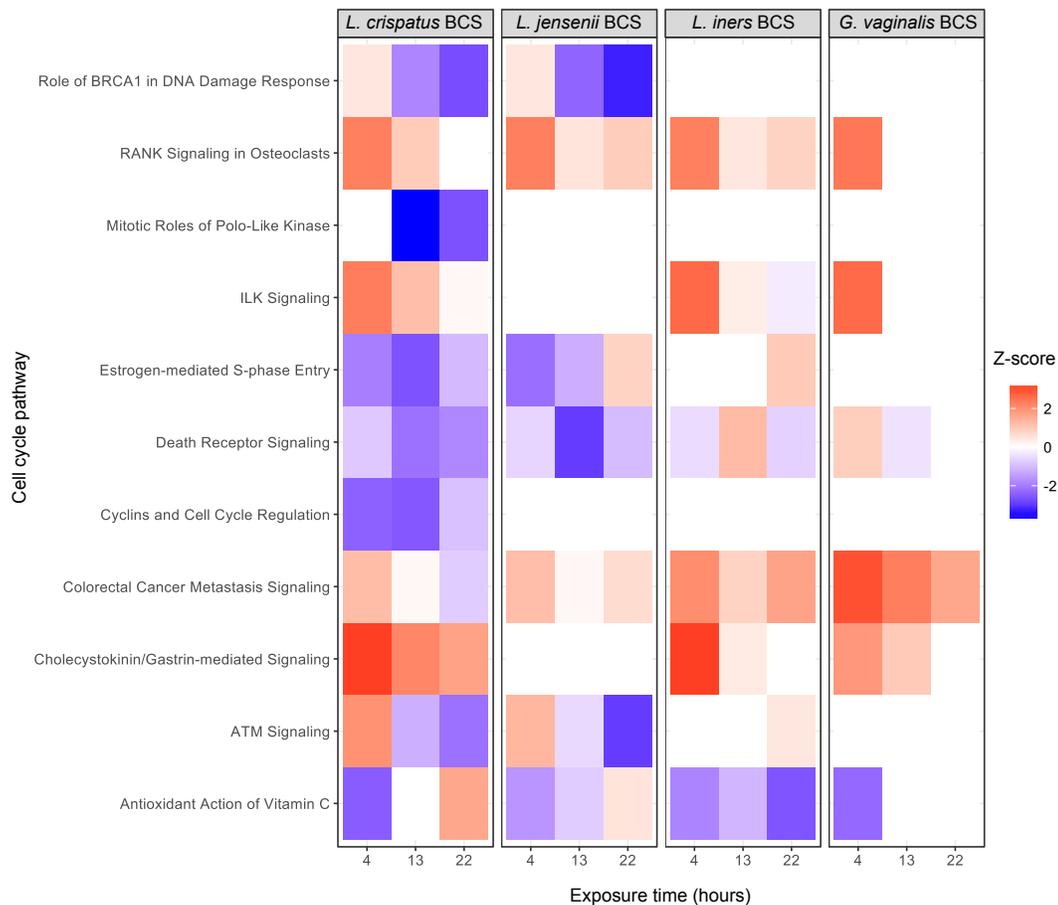


Figure 3.3 Heat map of activation scores (z-score) from pathways associated with the cell cycle for cells exposed to each BCS for 4h, 13h or 22h vs. cell culture medium.

Pathways with absolute z-score greater than 2 in at least one comparison revealed repression of cell cycle pathways after exposure to *L. crispatus* and *L. jensenii* BCS. Conversely, *L. iners* or *G. vaginalis* BCS are ineffective at repressing cell cycle pathways.

3.5 shows the IL-6 Signaling canonical pathway from IPA for *L. crispatus* BCS vs. cell culture medium at 13h (z-score 1.3). Note that upstream effectors are mostly decreased in *L. crispatus*, including IL-1R, the receptor for the pro-inflammatory cytokine IL-1 (green colors), but increased in downstream (red colors). These expression patterns negate the activation of the IL-6 Signaling Pathway and instead may indicate downstream genes are activated by a separate pathway other than IL-6 (Figure 3.4, Figure 3.5). Taken together, this suggests that *L iners* BCS maintains a

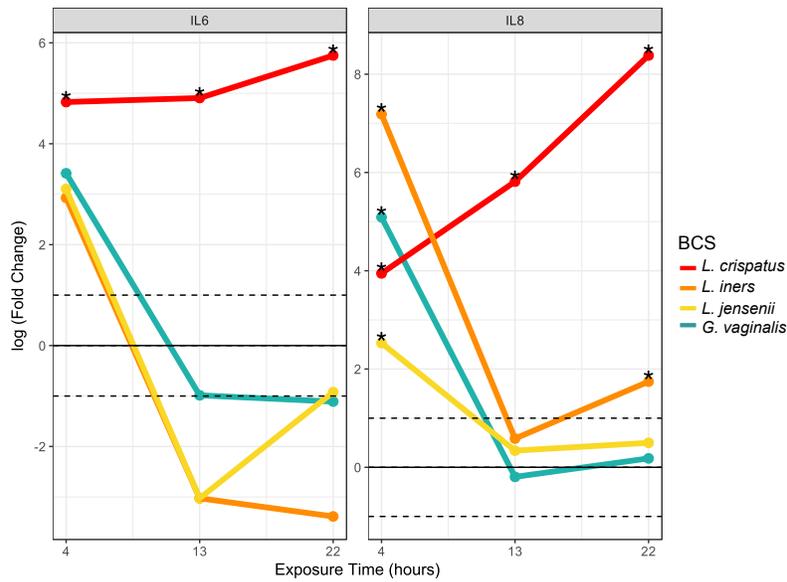


Figure 3.4 Longitudinal relative expression patterns for IL6 and IL8.

Shown are gene's log₂ fold change values for 4h, 13h, and 22h BCS exposure vs. cell culture medium. Line color indicates BCS exposure. Horizontal dotted lines within each plot indicate log₂ fold change -1, 0 and 1 and "*" above each BCS exposure time indicates FDR < 0.01.

relatively high stimulation of pro-inflammatory pathways over the time course, similar to *G. vaginalis*, while both *L. crispatus* and *L. jensenii* BCS initially stimulate pro-inflammatory pathways at 4h after BCS exposure. By 13h of exposure, *L. crispatus* BCS decreases host cell cycle pathways while simultaneously minimally inducing pro-inflammatory pathways (no pro-inflammatory pathways with z-score >2), and *L. jensenii* BCS primarily decreases host cell cycle with little effect on immune responses.

Histone modification and cell cycle regulators are expressed dependent on BCS

The cell cycle may be regulated by a global mechanism, given that multiple cell proliferation pathways were repressed in *L. crispatus* and *L. jensenii* BCS exposed cells. Figure 3.6 shows the Cyclins and Cell Cycle Regulation canonical

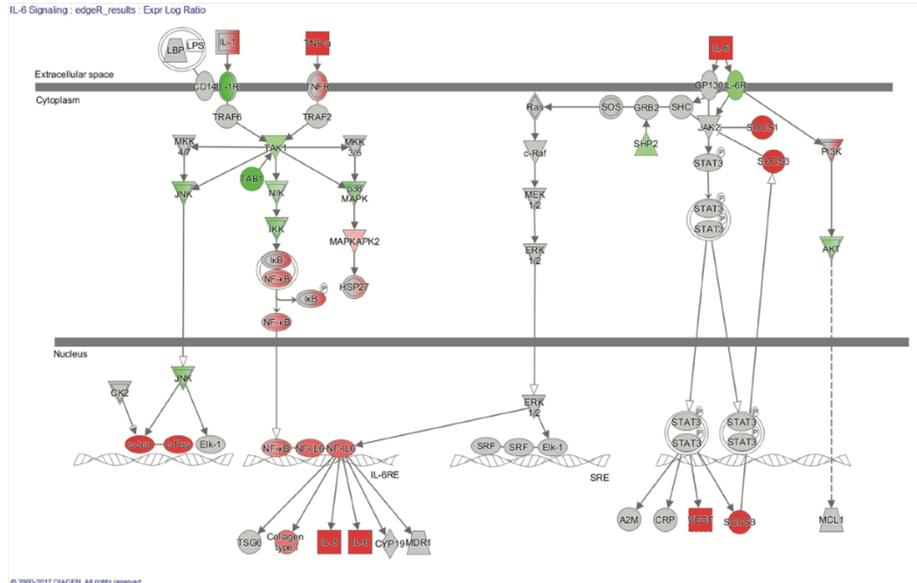
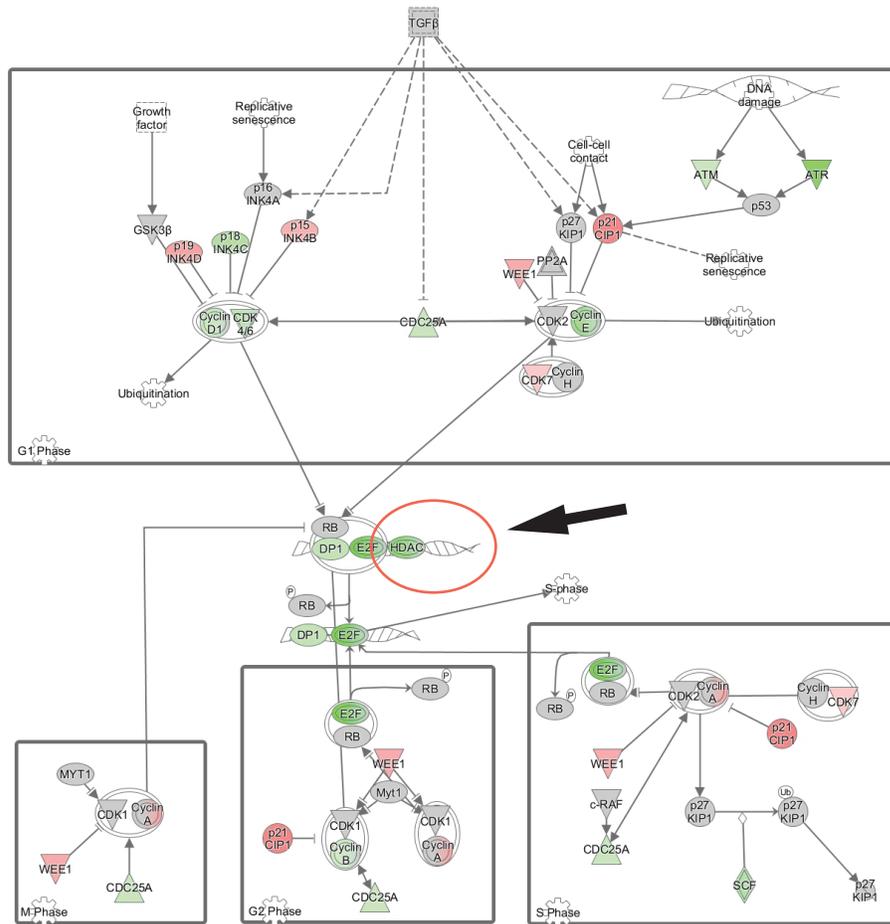


Figure 3.5 IL6/IL1 signaling pathways gene expression after exposure to *L. crispatus* BCS for 13h compared to cell culture medium.

Upstream differentially expressed genes show mostly decreased (green) logFC (IL-1), while downstream differentially expressed genes show increased (red) logFC, resulting in z-score 1.3. Pathway generated in IPA.

pathway from IPA. Differentially expressed genes within the pathway are colored by the mean logFC of both *L. crispatus* and *L. jensenii* comparisons at 13h after BCS exposure. Key transcriptional regulators of this pathway are histone deacetylase enzymes (HDACs, circled on Figure 3.6) which regulate gene transcription by chromatin remodeling [328]. This study found that the longitudinal expression of HDAC4 has a decreased logFC pattern at 13h after *L. crispatus* or *L. jensenii* BCS exposure (Figure 3.7). Interestingly, the histone acetyltransferase enzyme (HAT) E1A binding protein p300 (EP300, known to oppose the action of HDACs) exhibits the opposite longitudinal expression pattern of HDAC4 logFC at 13h (Figure 3.7).

It has previously been shown that HDAC4 negatively regulates the transcription of cyclin dependent kinase inhibitor 1A (CDKN1A) [329], which in turn



© 2000-2017 QIAGEN. All rights reserved.

Figure 3.6 Cyclins and Cell Cycle Regulation pathway for averaged logFC of *L. crispatus* BCS and *L. jensenii* BCS 13h after exposure.

Genes in pathway colored by differentially expressed genes with logFC > 0 (red) or log FC < 0 (green). Encircled is HDAC, which act in chromatin remodeling and global transcriptomic regulation of cell cycle. Pathway generated by IPA.

inhibits the activity of cyclin dependent kinase 4 (CDK4) and thus the cell cycle.

Figure 3.7 shows that the CDKN1A longitudinal logFC expression is greatest in *L. crispatus* BCS exposed cells at 13h, followed by *L. jensenii* BCS exposed cells, with *L. iners* and *G. vaginalis* exposed cells having logFC < 1 at 13h after BCS exposure. CDK4, on the other hand, shows logFC is most negative at 13h in *L. crispatus* BCS exposed cells, followed by *L. jensenii*, *L. iners* and finally *G. vaginalis* vs. cell culture medium (Figure 3.7). Chapter 2 showed that decreased protein expression of CCND1

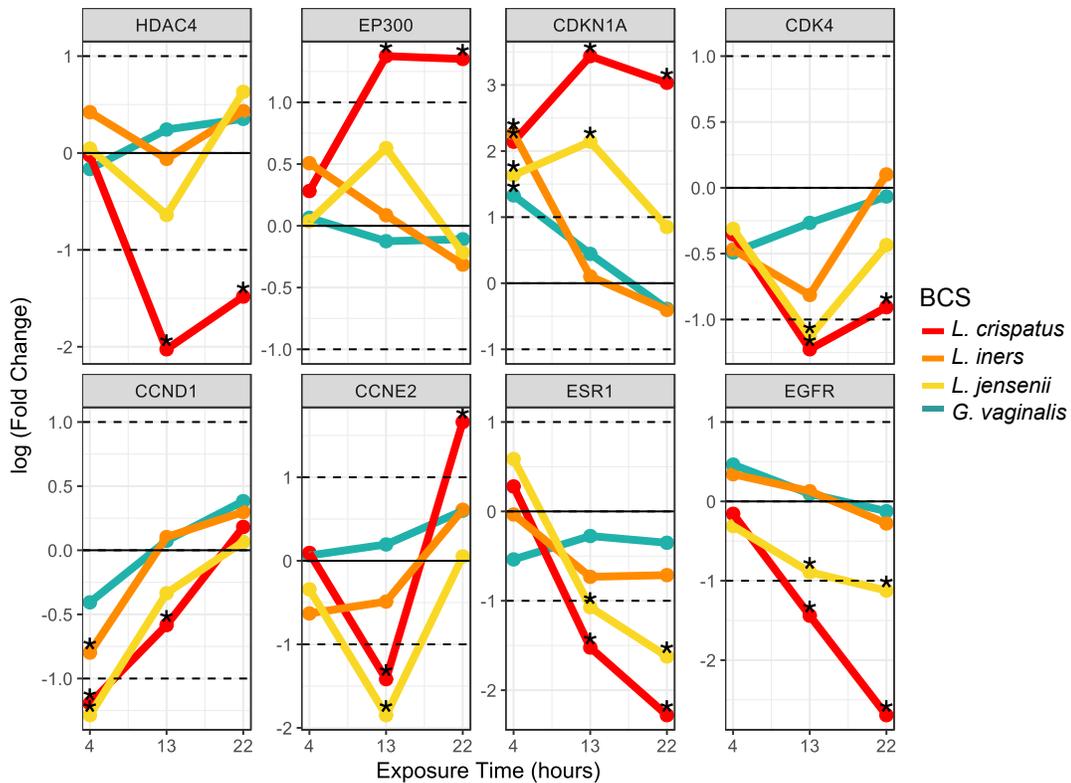


Figure 3.7 Longitudinal relative expression patterns for selected cell cycle and chromatin remodeling genes.

Shown are gene's log₂ fold change values for 4h, 13h, and 22h BCS exposure vs. cell culture medium. Line color indicates BCS exposure. Horizontal dotted lines within each plot indicate log₂ fold change -1, 0 and 1 and "*" above each BCS exposure time indicates FDR < 0.01.

occurs 13h after exposure to *L. crispatus* and *L. jensenii* BCSs. Figure 3.7 reveals that although CCND1 logFC increases from 4h to 22h over all comparisons relative to cell culture medium, the value is more negative for *L. crispatus* BCS, followed by *L. jensenii* BCS and finally *L. iners* BCS and *G. vaginalis* BCS. Cell cyclin E2 (CCNE2) is a regulator of late G1/S phase cell progression and shows decreased expression vs. cell culture medium in *L. crispatus* BCS and *L. jensenii* BCS at 13h relative to *L. iners* BCS or *G. vaginalis* BCS comparisons (Figure 3.7). The opposing logFC expression profiles of CDKN1A versus CDK4, CCND1 and CCNE2 at 13h after *L. crispatus* or *L. jensenii* BCS exposure are aligned with decreased cell

proliferation observed in Chapter 2 and that differential transcription of histone modification genes HDAC4 and EP300 further suggest global regulation of CDKN1A and the cell cycle.

Estrogen Receptor Alpha and Epidermal Growth Factor Receptor 1 gene expression are significantly decreased after *L. crispatus* and *L. jensenii* BCS exposure

Chapter 2 identified CCND1 as a target of miR-193b in cells exposed to *Lactobacillus* spp. BCS using experimentally validated targets from miRTarBase [284]. Another experimentally validated target of miR-193b is Estrogen Receptor Alpha (ESR1), which when expressed induces cell growth via estrogen [312], [313]. Figure 3.7 shows differential logFC expression (FDR <0.01) of ESR1 in *L. crispatus* and *L. jensenii* BCS exposed cells at 13h and 22h BCS exposure. This may be reflective of miRNA-193b-mediated regulation of ESR1 transcriptional expression in *L. crispatus* or *L. jensenii* BCS exposed cells.

Chapter 2 also found decreased cell proliferation was required for *C. trachomatis* infection. Epidermal Growth Factor Receptor 1 (EGFR) is required for internalization of elementary bodies into the host cell [316]. EGFR's logFC expression becomes more negative from 4h to 22h (Figure 3.7) with *L. crispatus* BCS and *L. jensenii* BCS exposed cells having FDR <0.01 at 13h and 22h. This suggests that decreased EGFR expression may be mediated by *Lactobacillus* spp. BCS exposure as an additional protective mechanism against *C. trachomatis* internalization.

Discussion

This chapter revealed global transcriptomics expression patterns in VK2 vaginal epithelial cells when exposed to *Lactobacillus* spp. or *G. vaginalis* BCS for 4h, 13h and 22h. The results described in this chapter show that both cell cycle and immune pathways were among the most activated or repressed pathways shared among all BCS and exposure times. In Chapter 2, *Lactobacillus* spp. BCS exposure decreased vaginal epithelial cell proliferation associated with increased expression of miR-193b resulting in reduced CCND1 protein expression, the target of miR-193b. Results from Chapter 3 suggest BCS-mediated cell cycle control is more broadly regulated by chromatin remodeling genes and other cell cycle regulators. On the other hand, exposure to *L. iners* and *G. vaginalis* BCS resulted mainly in immune pathway activation.

A family of enzymes that regulate the cell cycle are HDACs and HATs, which control gene expression by deacetylating or acetylating histones, respectively [328]. After exposure to *L. crispatus* and *L. jensenii* BCSs, HDAC4 expression is decreased while that of EP300 is increased. This is illustrated in Figure 3.6 and Figure 3.7 which depict the cell cycle pathway ‘Cyclins and Cell Cycle Regulation’ and the longitudinal expression profile of these genes. Histone modifications via HDAC regulation, which result in decreased access to DNA to transcriptional regulators, is known to be coupled with the immune response, for example through MAPK and NF- κ B signaling pathways, leading to increased or decreased IL-6 or IL-8 expression [330], [331]. Interestingly, bacterial-derived small chain fatty acid (SCFA), including butyrate, have been shown to activate HATs and inhibit HDACs in the

gastrointestinal tract [332], [333] in addition to modulating the immune response, for example via IL-8, IL-6, and IL-1 β [332], [334], [335]. Butyrate is known to be transported into the cell by solute carrier family 5 member 8 (SMCT1) [336] thus potentially having a more direct effect on intracellular proteins. Lactic acid, a SCFA produced in copious amounts by *Lactobacillus* spp., has been shown to stimulate the release of IL-1 β and IL-8 in poly(I:C) stimulated vaginal cells [102] and IL-23 [101] in PBMCs in a dose-dependent manner. Similar to butyrate, lactate can be transported across the cell through proton-linked monocarboxylate transporters in some cells such as rat or human intestinal epithelial cells (IEC-6 and Caco-2, respectively) [337]-[339]. The probiotic species *Lactobacillus plantarum* reverses decreased SMCT1 expression induced by TNF- α , suggesting that *Lactobacillus* may play an active role in butyrate uptake in the gut [337]. An unanswered question is whether lactate is transported into cells through a similar mechanism in the vaginal environment and whether its potential effect on HDACs or immune function has a similar role.

HDAC4 inhibition reduces cell proliferation by inducing transcription of the CDK4 inhibitor CDKN1A [340] [341] [329]. In line with this data, CDKN1A logFC expression is greatest at 13h post-exposure to *L. crispatus* and *L. jensenii* BCSs (Figure 3.7). Chapter 2 showed that the amount of CCND1 protein is decreased after 13h exposure to *L. crispatus* and *L. jensenii* BCSs. As miRNA mechanisms can also include transcriptional degradation, decreased CCND1 protein and transcript abundance is supported by results from Chapter 2 and Chapter 3. The findings presented here further suggest that decreased cell proliferation is mediated by HDAC4 de-repression of CDKN1A transcript expression.

The HAT EP300 regulates ESR1 by mediating ligand binding to the estrogen receptor [342]. The results presented here suggest an inverse relationship between EP300 expression and ESR1 transcripts (Figure 3.7). Estrogen is a mediator of maturation and cell growth of the vaginal epithelium [312], [313]. Thus, the observed inhibition of cell proliferation via *Lactobacillus* spp. BCS exposure in Chapter 2 and the known regulation of ESR1 by EP300 suggest a mechanism to maintain cell homeostasis. As this is a potentially complex feedback mechanism, the interpretation of this data in the context of the vaginal microbiota warrants further investigation.

A significant finding from Chapter 2 is that decreased cell proliferation decreases the efficiency of *C. trachomatis* infection *in vitro*. Key to the internalization of *C. trachomatis* elementary bodies into the host target epithelial cells is their attachment to EGFR [316]. Interestingly, EGFR expression is decreased over time after exposure to *L. crispatus* and *L. jensenii* BCS (Figure 3.7). Control of EGFR by these two *Lactobacillus* species may contribute to host increased resistance to *C. trachomatis* when *L. crispatus* and *L. jensenii* dominate the vaginal ecosystem. Here we show that EGFR expression is still occurring after exposure to *L. iners* and *G. vaginalis* BCS, which may result in a lack of resistance to *C. trachomatis* infection after exposure to these species. (Figure 3.7). Further research is needed to better understand the interaction between EGFR expression, vaginal microbiota and *C. trachomatis* infectivity.

Previous studies have found that relative to cell culture medium control, cytokine IL-1 β or IL-8 secretion is reduced in *L. crispatus* and *L. jensenii* exposed cells [59]. Instead, exposure to *L. iners* induces the pro-inflammatory cytokine tumor

necrosis factor (TNF), the cytokine signaling gene interleukin 1 receptor associated kinase 2 (IRAK2) and the pro-inflammatory transcriptional regulators interferon regulatory factor 1 (IRF1) and nuclear factor kappa B inhibitor alpha (NFKBIA) [59], [67], [87]. The BV-associated species *G. vaginalis* induces IL-6 secretion [87]. *L. crispatus* BCS was found to activate IL-6 and IL-8 pathways relative to cell culture medium 4h post-exposure (Figure 3.2). An explanation for inconsistencies between the reported literature and these results may be a disconnect between cellular transcriptional measurement performed here and the protein level measurement reported in the literature. In addition, the *in vitro* nature of the experimental methods might have an impact on the results. However, pathway-level analysis results are more aligned with expected reduced or activated immune responses. For example, IL-6 has a relatively low z-score of 1.3 13h after exposure to *L. crispatus* BCS. This is explained by the reduced expression of genes upstream from IL-6, such as IL-6 receptor (IL-1R), while IL-6 itself is slightly over-expressed. (Figure 3.2, Figure 3.4, and Figure 3.5). This observation suggests that the unexpected gene expression patterns may not result in activation of the pathway that ultimately determines epithelial cell response. Thus, although gene expression patterns are slightly unexpected, each BCS exposure aligns with previously reported cellular responses when pathway level analysis are performed.

Pathway activation suggests the *L. iners* BCS host response is more similar to that of *G. vaginalis* BCS than that of *L. crispatus* or *L. jensenii* BCS, even at later time points. Some strains of *L. iners* have been shown to induce inflammatory responses and encode inerolysin, a cholesterol dependent cytolysin (CDC), which is

known to activate pro-inflammatory responses and is similar to vaginolysin, a CDC produced by *G. vaginalis* [42][112] [112]. Additionally, *L. iners* only produces the L(+) lactic acid isomer [311]. In one study, acidic L(+)-lactic acid with poly(I:C) TRL3 stimulation produced IL-6 and IL-8 while the racemic lactic acid only produced IL1RA [100]. Therefore, there may also be an isomer-specific component wherein L(+)-lactic acid produces different responses than D(-) or D/L-lactic acid. These findings highlight the need for strain-specific studies in microbiome research.

Concluding remarks

This chapter aimed to characterize the transcriptomic response of VK2 vaginal epithelial cells to *Lactobacillus* spp. and *G. vaginalis* BCS, the latter as a surrogate to dysbiotic microbiota. Immune-related pathways including IL-6 and IL-8 were mostly activated in *L. iners* or *G. vaginalis* exposed BCS cells while cell proliferation pathways were inhibited by *L. crispatus* and *L. jensenii* at 13h after BCS exposure. Of note, immune functions have been linked to cell proliferation in other cells, and the results presented here posit that decreased cell proliferation upon *Lactobacillus* spp. exposure may be partially mediated by HDAC or lactate-specific effects via shared pathways. EGFR and ESR1 may be additionally regulated by *L. crispatus* and *L. jensenii* mediated exposed mechanisms such as HDAC inhibition. There is further need to understand strain or species specific effects as this may have profound consequences on vaginal health by modulating vaginal cell homeostasis.

Chapter 4 Future work and conclusions

Future Work

Studying the microbiota-host dynamic in more complex *in vitro* cell systems

The effects of the vaginal microbiota were studied using bacteria culture supernatants containing lactic acid. Although these BCS provided valuable insights, adding viable bacteria to the mammalian cells would better approximate the vaginal environment. However, this task is technically challenging as optimal bacterial and mammalian growth media and conditions differ. For example, many of the BV-associated microbes are difficult to culture, while others do not grow optimally in VK2s medium [343]. While it might be desirable to include bacterial cells in the system, in the case of *Lactobacillus* spp., it might not represent their interactions with the vaginal epithelium. Indeed, *Lactobacillus* spp. do not adhere to the vaginal epithelium and interact with the host epithelium through the panoply of metabolites and small molecules they are known to produce. Further, the system lacks epithelial shedding, which occurs *in vivo*, but is not observed *in vitro*, highlighting the limitation of working with cell cultures. However, the lack of animal models that mimic humans to study the vaginal environment is still the main limitation. *In vitro* findings need to be validated by measuring them directly in human from samples prospectively collected.

Chapter 2 and Chapter 3 findings suggest species-specific host responses. Many factors can explain these differences, first of which are genomic differences between species and strains of *Lactobacillus* or between strains of *G. vaginalis* [277].

Lactobacillus spp. are known to produce different amounts and ratios of D(-) and L(+)-lactic acid [104]. While Figure 2.6 did not show any significant differences between isomers in miR-193b expression, it might be involved in the regulation of other pathways through HDAC, for example. In addition to lactic acid isomer production differences, *L. iners* carries inerolysin, a cytolyisin whose function is still unknown but believed to be involved in innate immune activation [42]. Thus, for *L. iners*, carriage of inerolysin might have significant impact on vaginal homeostasis, and experiments using differing strains within the species, isolated from a woman with and without BV, would therefore be valuable to compare and further explain the findings presented here. Additionally, *G. vaginalis* was used as the prototypical BV-associated species. Experiments using other BV-associated bacteria should yield further insight into differences contributed by *Lactobacillus* spp. to the host homeostasis.

Using a three-dimensional cell system instead of a VK2 vaginal epithelial cell monolayer may better mimic *in vivo* processes. The vaginal epithelium is a polarized squamous stratified epithelium such that there is an apical and basal side, where cells become flattened and keratinized and shed when reaching the apical surface [5], [344]. With transwells, organotypic or organ-on-a-chip systems, it is possible to study the effect of basal vs. apical surface changes including basal estrogen or nutrient supply (as would be provided by the vasculature underlying the basal lamina in the vagina), microbial stimulation, and immune cells (such as macrophages) [345]. Using these models may allow the use of physiological pH for prolonged periods of time to better study the effect of reduced pH on the apical surface. Organ-on-a-chip systems

have demonstrated the requirement of peristalsis for replication of gut cell morphology as seen *in vivo* [346]. Adding movements that mimic the vaginal physical dynamic may provide additional improvements currently unknown to the model.

Further characterization of the role of miR-193b and mechanism of action

Results from Chapter 2 shows how *Lactobacillus* spp. exposure triggers miR-193b expression, which is known to directly target CCND1, so that after exposure to *L. crispatus* and *L. jensenii* BCS the amount of CCND1 is decreased. To further characterize the role of miR-193b, it would be valuable to manipulate the effect of miR-193b by increasing or decreasing its expression. This can be achieved by transfecting cells with a construct from which miR-193b can be overexpressed, allowing its expression in *in vitro* condition where it is usually repressed (exposure to *G. vaginalis* BCS or *L. iners* BCS at a certain time after exposure). Alternatively, interfering RNAs, RNA complementary to its target, would decrease available miRNA-193b and reverse the effects, confirming its role in the identified function [347]. There are several technical challenges to cell transfection, including the ability of the cell line to be transfected or infected by transducing viruses carrying the genetic constructs, unintended immune responses of the transfected cells or unknown genome mutagenesis associated with the process (as in the case of virus-mediated transduction) [348]. However, transfection of plasmids using cationic-lipid based methods has demonstrated that VK2s are receptive to DNA transfection [349] [350]. Cationic-lipid based methods are advantageous in that they have low cytotoxicity and

not associated with genome mutagenesis [348]. Transfection reporter assays are also advantageous for studying the effect of varying miRNA target sequence to protein expression efficiency. A dual luciferase reporter assay in which the luciferase amino acid would be coupled to CCND1 amino acid could demonstrate miR-193b direct role in CCND1 translational inhibition [351].

Evaluate the effect of HDAC inhibitors on cell proliferation, miR-193b expression and *C. trachomatis* infectivity

Chapter 3 found that HDAC4 is decreased in *Lactobacillus* spp. BCS cells. HDACs are known to have an active role in increasing cell proliferation and broad regulatory changes through histone modifications [352]. HDAC inhibitors are a class of commercially available molecules used primarily in cancer therapy to reduce cell proliferation [353]. Thus, a key experiment would be to test the effect of HDAC inhibitors on cell proliferation in VK2s while simultaneously measuring miR-193b expression, CCND1 expression and *C. trachomatis* infectivity. This may allow repurposing of a safe HDAC inhibitor for approved use as a microbicide after exposure to *C. trachomatis*, fostering a novel approach to STI treatment and prevention.

Elucidate the regulatory pathways leading to miR-193b expression

Transcription factors that govern the expression of miRNAs are largely unknown [146]. miR-193b may be transcribed along with other key regulatory mRNAs including regulators of immune pathways. Identifying miR-193b transcriptional regulators will help further link vaginal microbiota composition and

cell proliferation. This can be accomplished using chromatin studies including CHIP-seq in which all binding of a given transcription factor is sequenced.

Conclusions

This dissertation aimed to characterize miRNA and miRNA-associated mRNA regulatory pathways as a function of vaginal microbiota community composition differences. Chapter 2 applied Random Forest models to small RNA-seq expression profiles from 100 *Lactobacillus* spp.-dominated and CST-IV microbiota vaginal samples and uncovered miRNAs targeting functions such as transcription regulation, cell cycle, signaling, hypoxia, development and the immune response. *In vitro* experiments confirm that miRNA-193b expression was induced in vaginal epithelial cells exposed to *Lactobacillus* spp. bacterial culture supernatants, which was simultaneously shown to decrease cell proliferation and coincide with decreased CCND1 protein expression. Decreased *C. trachomatis* infectivity was observed because of decreased cell proliferation (Figure 4.1). These findings have profound implications for women's health as having a *Lactobacillus* spp.-dominated microbiota is critical for the reduction of cell proliferation and therefore *C. trachomatis* infectivity, thus raising important questions about the implications of not having *Lactobacillus* spp.-dominated microbiota. One can characterize non-*Lactobacillus* spp.-dominated microbiota as a normal state (as it is found in many women who are asymptomatic) that carry risk (as it is associated with increased risk of STIs). From a clinical point of view, it might be necessary to redefine the treatment guidelines which only recommend treatment if CST-IV is associated with reported symptoms.

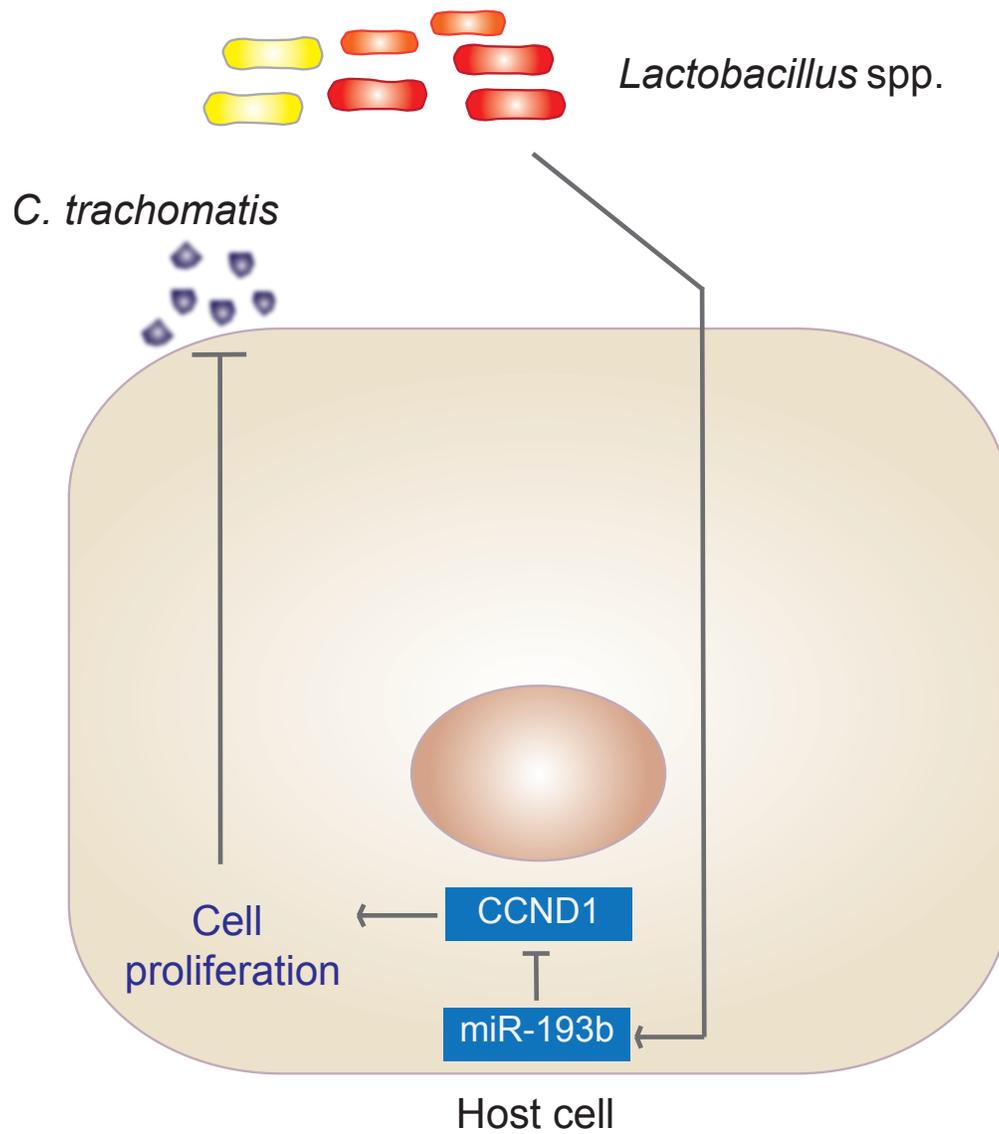


Figure 4.1 Summary of Chapter 2 findings

Chapter 2 demonstrated that a *Lactobacillus-spp.* dominated CST or BCS induces expression of miR-193b, which in turn decreases CCND1, a regulator of cell cycle, and therefore results in decreased cell proliferation. A critical finding from Chapter 2 is that cell proliferation is required for *C. trachomatis* infectivity, and that conversely, inhibiting cell proliferation via *Lactobacillus spp.* exposure may reduce *C. trachomatis* infection.

There might be instances where treatment might be recommended even though no symptoms are reported. However, efficient treatments are still lacking, as current treatment with metronidazole is associated with high recurrence. This research might lead to novel treatments to modulate the microbiota but also improve preventive and curative measures for *C. trachomatis* infection or exposure.

Given that the miRNA regulatory cascade is most likely mediated by a global response to vaginal microbiota, Chapter 3 sought to characterize whole transcriptomic changes associated with *Lactobacillus* spp. and non-*Lactobacillus* BCS. The direction and magnitude that each BCS activated or repressed genes and pathways suggest that the effects of BCS exposure are mainly mediated through modulation of immune and cell cycle pathways (Figure 4.2). Interestingly, all three *Lactobacillus* supernatants showed at least mild activation of immune response pathways at 4h BCS exposure, but *L. iners* maintained an activation of immune genes and pathways 13h after BCS exposure. *L. crispatus* and *L. jensenii* BCS exposure repressed cell cycle pathways whereas *L. iners* and *G. vaginalis* BCS were not as effective. These results suggest *L. iners* induces a host response more similar to that of *G. vaginalis* than other *Lactobacillus* spp. In line with this finding, *L. iners* did not induce high miR-193b expression in Aim1 which coincided with increased CCND1 protein expression. This highlights the importance of considering species, strains or specific metabolites when studying vaginal microbial species.

The cell cycle is positively regulated in part by chromatin remodeling via HDAC4-mediated repression of CDKN1A [329]. In concordance with this, HDAC4 and CDKN1A have opposing differential mRNA expression in *L. crispatus* and

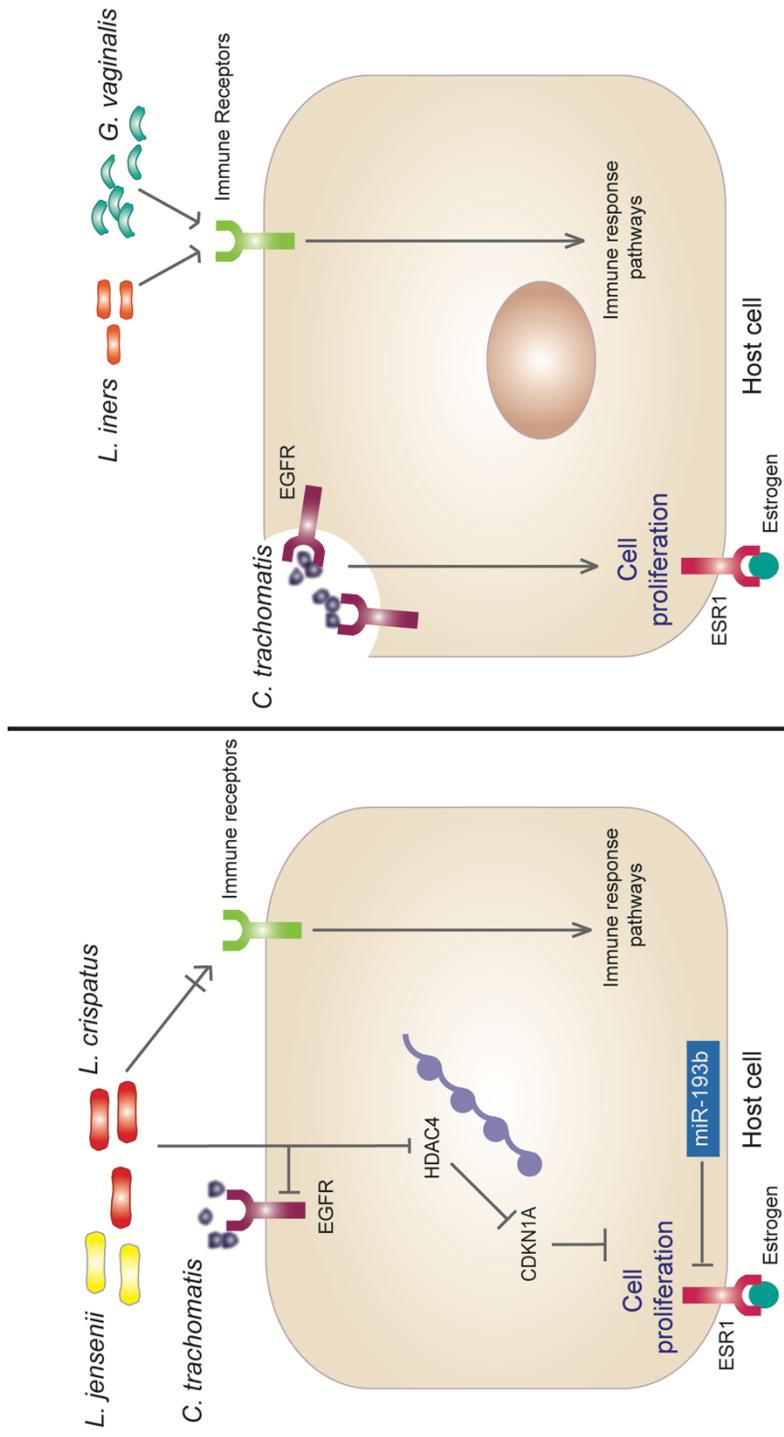


Figure 4.2 Summary of Chapter 3 findings

Chapter 3 demonstrated that in addition to reduced proliferation via CCND1, transcripts of other cell cycle regulators are also reflective of *L. crispatus* or *L. jensenii* BCS exposure, for example, reduced HDAC4 expression which is known to inhibit the cell cycle inhibitor CDKN1A. An experimentally validated target of miR-193b is Estrogen Receptor alpha (ESR1) whose transcription was decreased in *L. crispatus* or *L. jensenii* BCS exposed cells. As estrogen promotes vaginal cell growth and maturation, dampening of cell proliferation via *Lactobacillus* spp. may result in a homeostatic equilibrium. *L. iners* and *G. vaginalis* BCS exposure do not reduce cell proliferation pathways as effectively as *L. crispatus* or *L. jensenii* BCS, and thus estrogen is the driver of cell proliferation (i.e. shedding, as outlined in Figure 2.12). Other processes such as decreased EGFR availability, required for *C. trachomatis* infection, might be affected by vaginal microbiota. *L. iners* and *G. vaginalis* BCS were also found to trigger pro-inflammatory responses whereas this response was not as apparent in *L. crispatus* or *L. jensenii* BCS exposed cells.

L. jensenii BCS exposed cells, corresponding to the observed decreased cell cycle in Chapter 2. Furthermore, expression of ESR1, a target of miR-193b, was decreased in *L. crispatus* and *L. jensenii* BCS exposed cells, which may additionally desensitize cells to estrogen-mediated proliferation (Figure 4.2). An additional protective mechanism against *C. trachomatis* infection is supported by decreased EGFR transcript abundance mediated by of *L. crispatus* or *L. jensenii* BCS (Figure 4.2). Taken together, these findings suggest that through decreased epithelial cell proliferation and miRNA-mediated regulation after exposure to *L. crispatus*, *L. jensenii*, and to a lesser extent *L. iners*, the vaginal microbiota protects the vaginal epithelium against STI such as *C. trachomatis*. Decreasing cell proliferation can be achieved potentially through *Lactobacillus* probiotics, lactic acid, or directly through HDAC inhibitors. These results highlight the need to better understand the host response to microbiota, as potential therapeutics including lactate, probiotics or HDAC inhibitors could be repurposed for novel prevention or treatment methods of STIs and maintaining a woman's vaginal homeostasis.

Appendices

Appendix 1 Steven Smith's contributions to thesis

Chapter 1

- Subsection 1 is a published review in the Journal of Physiology (2017). I performed secondary research for this review, and Jacques and I wrote the manuscript (May 2015 – October 2015). I created Figure 1.4 as part of the review's abstract figure.
- I performed research for the content in the sections on miRNA biogenesis/mechanism, RNA challenges and bioinformatic/computational approaches. I created all remaining figures in this chapter (Figure 1.3, Figure 1.5, Figure 1.6).

Chapter 2

- I designed and executed the project, experiments and analyses, and wrote the chapter that is currently under review as a manuscript submitted to mBio (2017). Specifically:
 - o I optimized the total RNA extraction protocol from archived vaginal swabs to achieve consistent yield and optimal RINe quality metrics (October 2013 – November 2014).
 - o I evaluated and identified a small RNA-seq library construction method to achieve consistent libraries containing reduced adapter-dimer formation (November 2013-April 2015)
 - o I extracted over 130 vaginal swab samples and constructed over 113 small RNA-seq libraries (May 2015- July 2016).

- I designed and built a bioinformatic pipeline to trim and QC reads, align reads to references, and count overlaps (January 2015 –April 2015).
- I designed, built and executed a computational pipeline to preform QC, and implement the Random Forest algorithm to analyze the small RNA-seq data (April 2015 – December 2015).
- I cultured vaginal bacteria involved in all bacterial culture supernatant experiments (April 2016 – April 2017).
- I cultured and maintained vaginal epithelial VK2 cells for *in vitro* experiments (April 2016 – April 2017).
- I designed, optimized and executed VK2 culture experiments including the VK2 BCS exposure time-course miR-qPCR, VK2 cell BCS exposure scratch/proliferation assay and VK2 cell lactic acid exposure miR-qPCR/scratch assay (May 2016 – April 2017).
- I designed, optimized and executed the BCS exposed VK2 Western blot for CCND1 and β -actin (December 2016 – March 2017).
- I generated all figures and tables within the chapter.
- I extracted approximately 2,000 vaginal swabs for 16S rRNA gene sequencing, with assistance from Bilal Iqbal and Latey Bradford (June 2013 – September 2016).
- I executed Pawel Gajer and Johanna Holm’s “Pecan”/MCMC-based 16S rRNA metataxonomic pipeline to classify reads to taxonomy (May 2016 – October 2016).

- Parts where I had assistance:
 - The Institute for Genome Science Genome Resource Center sequenced small RNA-seq libraries.
 - Pawel Gajer initially created the “subject-specific” Random Forest script. I modified it to include permuted Random Forest implementation.
 - Vonetta Edwards designed and executed *C. trachomatis* infection experiments and analyses. She also edited the manuscript.
 - Michael Humphrys tracked and coordinated 16S rRNA sequencing samples.
 - Larry Forney and Patrik Bavoil contributed to interpretation of the data and edited the manuscript.
 - Jacques conceived the project and experiments and helped me analyze and interpret the data. He also helped write the manuscript and thesis chapter.

Chapter 3

- I designed and executed the project, experiments and analyses, and wrote the chapter that will be shortly converted into a manuscript. Specifically:
 - I cultured vaginal bacteria involved in all bacterial culture supernatant experiments (November 2016 – February 2017).
 - I grew VK2 epithelial cells for 4h, 13h and 22h in triplicate exposed to BCS (November 2016 – February 2017).

- I extracted total RNA from 45 samples of VK2 cells exposure to BCS (February 2017).
- I constructed 44 rRNA-reduced RNA-seq libraries (February 2017).
- I designed, built and executed a bioinformatic pipeline to trim and QC the reads, align to reference and count overlaps (April 2015 – December 2015).
- I designed, built and executed a computational pipeline to perform QC, differential expression and pathway enrichment analysis (March 2017).
- I generated all figures and tables within the chapter.
- Parts where I had assistance:
 - The Institute for Genome Science Genome Resource Center sequenced rRNA-reduced RNA-seq libraries.
 - Jacques conceived the project and experiments and helped me analyze and interpret the data. He also helped write thesis chapter.

Chapter 4

- I performed research for the content in section future work and created all figures.

Appendix 2 Scripts for Chapter 2 & 3 Analysis (R markdown file and

rfSubjectSpecific.R)

```
##-----
## rfSubjectSpecific.R
## This script runs RF k-fold cross validation on the subject level
## Subject-specific implementation written by Pawel Gajer, updated 7/13/16, and
## adapted and updated by Steven Smith, 4/4/17
##-----

library(randomForest)
library(parallel)
library(rfPermute)

rfSubjectSpecific <- function(X, y, subjID, nfolds=10, verbose=FALSE,nrep=100,
permute=TRUE,... )

  ## Arguments:
  ## X    - predictors; data frame or matrix
  ## y    - response; vector of the same length as nrow(X)
  ## subjID - vector of length nrow(X) with subject assignment to each sample (used
in )
  ## nfolds - number of folds
  ## nrep   - number of permutations (ignored if permute=FALSE)
  ## permute- whether rfPermute() should be used instead of randomForest()
  ## verbose -print progress/output
  ## ... - parameters to pass to randomForest() or rfPermute()

  ## Values:
  ## - error: list of differences between prediction and true values for
  ## regression and pred==true logical vectors for classification (one for each
  ## fold)
  ## - rmse: root mean squared error
  ## - nmse: normalized MSE = MSE/MSE(mean(y) as predictor) for regression and
  ## MSE/MSE(the highest frequency class as predictor ) for classification
  ## - mae: mean absolute error
  ## - nmae: normalized MAE
  ## - cl.err: classification error = accuracy = sum(pred!=y)/length(y)

  {
    if(!is.numeric(y)){ ## Classification has 2 more importance metrics that regression
does, plus one for each class
      nMetrics<-2+length(unique(y))
    }else{ ## regression only has 2 RF metrics: IncNodPurity & IncMSE
```

```

nMetrics<-2
}
n <- length(y)

if ( !is.numeric(y) && !is.factor(y) )
  stop("y should be either numeric or factor")

if ( !is.data.frame(X) && !is.matrix(X) )
  stop("X is neither data frame nor matrix")

if ( n!= nrow(X) )
  stop("n!= nrow(X)")

if ( n!= length(subjID) )
  stop("n!= length(subjID)")

if ( !is.numeric(nfolds) )
  stop("nfolds is not numeric")

if ( nfolds < 1 )
  stop("nfolds < 1")

if ( nfolds > n )
  stop("nfolds > n")

if ( length(y[is.na(y)]) > 0 )
{
  warning("y has some NAs; removing them and the corresponding rows of X and
subjID")
  idx <- !is.na(y)
  y <- y[idx]
  X <- X[idx,]
  subjID <- subjID[idx]
}

subjID <- as.character(subjID)
uqSubjIDs <- unique(subjID)
nSubj <- length(uqSubjIDs)

if ( nfolds > nSubj )
{
  warning(paste("nfolds needs to be not greater than number of subjects. Changing it
to number of subjects: ",nSubj, sep=""))
  nfolds <- nSubj;
}

```

```

sp <- split(1:n, list(factor(subjID))) # list with each entry being a
# vector of indices of y that
# correspond to the same subjID
# (used as the label of the list
# element)
y.mean <- 0
y.mostFr <- 0
if ( is.numeric(y) )
{
  y.mean <- mean(y)
}

### # The splitting of data is done on the subject level. In particular, if
### # nfolds is equal to the number of subjects, we get a jack-knife
### # leave-one-out CV.
### # Note that y does not have to be constant over subjects.

s0 <- split(sample(nSubj),rep(1:nfolds,length=nSubj)) # nfolds split of all subjects
## Turning each element of s0 from vector of subject indices to vector of
## sample indices corresponding to the given subjects
s <- list()
for ( i in seq(s0) )
{
  v <- uqSubjIDs[s0[[i]]]
  s[[i]] <- as.vector(unlist(sp[v]))
}

x.null <- rep(y.mostFr, n)

error <- list() # list of prediction errors: prediction - y
error.null <- list() # list of null model prediction errors: mean(y) - y;
imp_permute<-list() ## container for importance metrics for rfPermute
mdl<-list()
# mean(y) is predicted value for each coordinate - null
# model; I am returning error.null so we can test if the
# current model is significantly better than the null
# model - that is if the mean(abs(errors)) is
# significantly different from the
# mean(abs(null.errors))

sampleIdx <- c() # vector of sample indices from each run of cross validation, so
we can match errors with samples
r2.loc <- numeric(nfolds)
r2.pearson.loc <- numeric(nfolds)
r2.spearman.loc <- numeric(nfolds)

```

```

rmse.loc <- numeric(nfolds)
nmse.loc <- numeric(nfolds)
mae.loc <- numeric(nfolds)
nmae.loc <- numeric(nfolds)
cl.err.loc <- numeric(nfolds)
ncl.err.loc <- numeric(nfolds) # normalized classification error
gError <- numeric(n) # "global" error array whose i-th entry is 1 if in the 10 fold CV
the prediction of y[i] was correct
imp <- matrix(0, nrow=ncol(X), ncol=nMetrics)
importance_w_pval <- matrix(0, nrow=ncol(X), ncol=2*nMetrics) ##stores P vals
and metrics, used in permutation
for (i in seq(nfolds))
{
  #i<-1
  if ( verbose )
    print(paste(" i=",i, sep=""))
  sampleIdx <- c(sampleIdx, s[[i]])
  trIdx <- setdiff(1:n, s[[i]])
  if(permute){
    print(paste0("Running CV fold ",i," out of ",nfolds ," using rfPermute"))
    m.rf <- rfPermute( X[trIdx,], y[trIdx], importance=TRUE,nrep = nrep, ... ) ##
returns rfPErmute object, which is radnomForest object with additional results
    importance.i<-rp.importance(m.rf) ## includes p values calculated from
rermutation model
  }else{
    print(paste0("Running CV fold ",i," out of ",nfolds ," using randomForest
(no NULL distirbutions will be generated)))
    m.rf <- randomForest( X[trIdx,], y[trIdx], importance=TRUE, ... )
    importance.i<-NULL
  }
  model.i<-m.rf ## saves the rf model object for iteration i

  ##m.rf <- randomForest( X, y, subset=setdiff(1:n, s[[i]]), importance=TRUE, ... )
  ##m.rf <- randomForest( X, y, subset=setdiff(1:n, s[[i]]), importance=TRUE)
  x <- predict(m.rf, newdata=X[s[[i]],], type="response")

  if ( is.numeric(y) ) ## If regression
  {
    error[[i]] <- x - y[s[[i]]]
    error.null[[i]] <- y.mean - y[s[[i]]]
    rmse.loc[i] <- sqrt(mean( error[[i]]^2 ))
    r2.loc[i] <- 100 * ( 1 - sum( error[[i]]^2 ) / sum( (x - y.mean)^2 ) ) # percentage of
variance explained
    r2.pearson.loc[i] <- 100*cor(x, y[s[[i]])^2
    r2.spearman.loc[i] <- 100*cor(x, y[s[[i]]], method="spearman")^2
    nmse.loc[i] <- mean( error[[i]]^2 ) / mean( (x - y.mean)^2 )

```

```

mae.loc[i] <- mean( abs( error[[i]] ) )
nmae.loc[i] <- mean( abs( error[[i]] ) ) / mean( abs(x - y.mean) )
imp <- imp + importance(m.rf)

} else { ## If classification
  m <- length(s[[i]])
  error[[i]] <- as.character(x) != as.character(y[s[[i]])]
  error.null[[i]] <- as.character(x.null[1:m]) != as.character(y[s[[i]])]
  cl.err.loc[i] <- sum(error[[i]]) / m
  ncl.err.loc[i] <- cl.err.loc[i] / ( sum(error.null[[i]]) / m ) # NOTE that this will be
NaN when the denominator is 0 (null model has no errors for the given y[s[[i]])
  ##print(cbind(as.character(x),as.character(y[s[[i]])]))
  imp <- imp + importance(m.rf)[,3:4] ##why were only the last 2 being used?
  #head(importance(m.rf))
}
gError[s[[i]]] <- as.integer(error[[i]])
imp_permute[[i]]<-importance.i
mdl[[i]]<-model.i
if(permute){
  importance_w_pval<-importance_w_pval+rp.importance(m.rf)
  head(importance_w_pval)
}
}

list( #imp_permute=imp_permute,
      importance_w_pval=importance_w_pval/nfolds,
      mdl=mdl,
      error=error,
      error.null=error.null,
      sampleIdx=sampleIdx,
      imp=imp/nfolds,
      rmse=mean(rmse.loc), nmse=mean(nmse.loc),
      mae=mean(mae.loc), nmae=mean(nmae.loc),
      r2=mean(r2.loc),
      r2.pearson=mean(r2.pearson.loc),
      r2.spearman=mean(r2.spearman.loc),
      gError=gError,
      cl.err=cl.err.loc,
      mean.cl.err=mean(cl.err.loc),
      ncl.err=ncl.err.loc,
      mean.ncl.err=mean(ncl.err.loc)
)
}

```

```

## normalized accuracy and classification error for a two-class classification
## problem

## predVals <- pls.sPTB.v2.predict
## trueVals <- sptb2.char.f

normClErr <- function(predVals, trueVals)
  ## predVals - predicted values
  ## trueVals - true values
  {
    if ( length(trueVals) != length(predVals) )
      {
        stop("ERROR: length(trueVals) != length(predVals)")
      }

    if ( !is.factor(trueVals) )
      {
        trueVals <- factor(trueVals)
      }

    predVals <- factor(predVals, levels=levels(trueVals))

    nElts <- length(trueVals)

    ## confusion matrix
    cm <- table(predVals, trueVals)

    ## accuracy
    acc <- (cm[1,1] + cm[2,2])/ nElts

    ## classification error
    clErr <- (cm[1,2] + cm[2,1])/ nElts

    ## accuracy of the naive classifier
    tt <- table(trueVals)
    i.mostFr <- which.max(tt)[[1]]
    true.mostFr <- names(tt)[i.mostFr]

    pred.null <- rep(true.mostFr, nElts)

    ## confusion matrix for the null model
    cm.null <- table(pred.null, trueVals)

    ## accuracy of the null model
    acc.null <- cm.null[1, i.mostFr] / nElts

```

```

## classification error of the null model
if ( i.mostFr == 1 )
{
  clErr.null <- cm.null[1,2]/ nElts
} else {
  clErr.null <- cm.null[1,1]/ nElts
}

## normalized accuracy
norm.acc <- acc / acc.null

## normalized classification error
norm.clErr <- clErr / clErr.null

print(paste("Accuracy:", acc))
print(paste("Accuracy of the naive classifier:", acc.null))
print(paste("Relative Accuracy:", norm.acc, " ## Should be way greater than 1"))
print(paste("Classification error:", clErr))
print(paste("Classification error of the naive classifier:", clErr.null))
print(paste("Relative Classification Error:", norm.clErr, " ## Should be very close to
0"))

invisible(list(cm=cm, cm.null=cm.null, acc=acc, acc.null=acc.null,
norm.acc=norm.acc, clErr=clErr, clErr.null=clErr.null, norm.clErr=norm.clErr))
}
##-----
## END rfSubjectSpecific.R
##-----

```

Thesis Analysis, R Markdown/Scripts

Steven Smith

May 4, 2017

Table of Contents

Analysis Script.....	137
Prepare Enviornment	137
Summary Statistics & Quality Control	161
Alignment Stats.....	161
Quality Control (QC) small RNA-seq count data.....	161
Discovery of miRNAs Associated w/ Lactobacillus spp. vs CST-IV/BV.....	182
Build a proxy-Amsel model from Clinical Visits	182
Normalize & log2 transform small RNA-Seq counts for use in models.....	188
Prepare & Analyze the Nugent and Amsel input tables (predictors) to RF model.....	196
Run Random Forest Models	198
Map miRNAs to Gene Targets and GO Process (miR-GO-Target) (Table 2).....	210
Function of miR-193b and in vitro Experimentation, Implications.....	219
Validate SmallRNASeq Results using qPCR.....	219
miR-193b qPCR Time-Course in VK2 monolayer after BCS Exposure	220
Quantify VK2 proliferation (function of miR-193b) exposed to BCS	224
Inhbit A2EN cell proliferation & Observe CT Infectivity	230
Subject Longitudinal Plots (Figure 1)	236
Ribo-reduced RNA-seq Analysis	245
Read counts data in from server	246
Create and Prepare Tables.....	247
Plot Replicates.....	254
Drop samples	254
Create model matrix	255
edgeR GLM FIT	255
Perfom edgeR LRT.....	256
Read in IPA Results	257
Extract and plot logFC values from LRT table.....	264

Analysis Script

The following markdown serves as a record for analysis performed in Steven Smith's PhD Thesis, 2017.

Change the root directory to point to where "Script_input" and "Scripts" directories reside.

This script will automatically create the following directories, if not present already:

1. Tables- Tables associated with thesis, in .csv or .txt
2. Figures - Figures associated with thesis, in .eps
3. Script_output - anything else output that is not a figure or table

A sessionsInfo log file will be included in Script_output.

Script_input contains static data used by script (previously generated and formatted for use as input). Also includes a data dictionary for sample metadata.

Some library installation may be required (e.g., `install.packages("PACKAGE")`)

```
timestamp()
## ##----- Wed May 31 23:50:58 2017 -----##
# Clear enviornment variables, set root
rm(list=ls())

## Set the root directory here:
root_directory<-"~/Dropbox
(IGS)/Jacques_Steve_Shared/Thesis/Thesis_pipeline/AnalysisPipeline/"

## Supress warnings to make knit PDF shorter... but turn these back on since
there may be some weird behaviors
knitr::opts_chunk$set(warning=FALSE, message=T,size=8)
```

Prepare Enviornment

1. Setup environment, variables, etc

2. Define custom functions

```
## Load libraries
## Note that some packages mask others. This might be a problem for, e.g.,
rename.

## Load libraries.
## Note that some override/mask functions from others. Had to explicitly use
```

```

the "dplyr" package for all "select" statements
library(reshape)
library(ggbiplot)

## Loading required package: ggplot2

## Loading required package: plyr

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:reshape':
##
##   rename, round_any

## Loading required package: scales

## Loading required package: grid

library(plyr)
library(grid)
library(scales)
library(gridExtra)
library(edgeR)

## Loading required package: limma

library(Biobase)

## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##   clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##   clusterExport, clusterMap, parApply, parCapply, parLapply,
##   parLapplyLB, parRapply, parSapply, parSapplyLB

## The following object is masked from 'package:limma':
##
##   plotMA

## The following object is masked from 'package:gridExtra':
##
##   combine

## The following objects are masked from 'package:stats':
##
##   IQR, mad, xtabs

```

```

## The following objects are masked from 'package:base':
##
##   anyDuplicated, append, as.data.frame, cbind, colnames,
##   do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##   grepl, intersect, is.unsorted, lapply, lengths, Map, mapply,
##   match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##   Position, rank, rbind, Reduce, rownames, sapply, setdiff,
##   sort, table, tapply, union, unique, unsplit, which, which.max,
##   which.min

## Welcome to Bioconductor
##
##   Vignettes contain introductory material; view with
##   'browseVignettes()'. To cite Bioconductor, see
##   'citation("Biobase)", and for packages 'citation("pkgname)".

library(RColorBrewer)
library(psych)

##
## Attaching package: 'psych'

## The following objects are masked from 'package:scales':
##
##   alpha, rescale

## The following objects are masked from 'package:ggplot2':
##
##   %+%, alpha

library(randomForest)

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:psych':
##
##   outlier

## The following object is masked from 'package:Biobase':
##
##   combine

## The following object is masked from 'package:BiocGenerics':
##
##   combine

```

```

## The following object is masked from 'package:gridExtra':
##
##   combine

## The following object is masked from 'package:ggplot2':
##
##   margin

library(rfPermute)
library(tidyverse)

## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Conflicts with tidy packages -----
-

## %+%():      ggplot2, psych
## alpha():    ggplot2, psych, scales
## arrange():  dplyr, plyr
## col_factor(): readr, scales
## combine():  dplyr, randomForest, Biobase, BiocGenerics, gridExtra
## compact():  purrr, plyr
## count():    dplyr, plyr
## discard():  purrr, scales
## expand():    tidyr, reshape
## failwith(): dplyr, plyr
## filter():   dplyr, stats
## id():       dplyr, plyr
## lag():      dplyr, stats
## margin():   ggplot2, randomForest
## mutate():   dplyr, plyr
## Position(): ggplot2, BiocGenerics, base
## rename():   dplyr, plyr, reshape
## summarise(): dplyr, plyr
## summarize(): dplyr, plyr

library(squash)
library(stringr)
library(plotly)

##
## Attaching package: 'plotly'

## The following objects are masked from 'package:plyr':
##
##   arrange, mutate, rename, summarise

```

```
## The following object is masked from 'package:ggplot2':  
##  
##   last_plot  
  
## The following object is masked from 'package:reshape':  
##  
##   rename  
  
## The following object is masked from 'package:stats':  
##  
##   filter  
  
## The following object is masked from 'package:graphics':  
##  
##   layout  
  
library(gPCA)  
library(nlme)  
  
##  
## Attaching package: 'nlme'  
  
## The following object is masked from 'package:dplyr':  
##  
##   collapse  
  
library(caret)  
  
## Loading required package: lattice  
  
##  
## Attaching package: 'caret'  
  
## The following object is masked from 'package:purrr':  
##  
##   lift  
  
## The following object is masked from 'package:rfPermute':  
##  
##   confusionMatrix  
  
library(gplots)  
  
##  
## Attaching package: 'gplots'  
  
## The following object is masked from 'package:squash':  
##  
##   bluered  
  
## The following object is masked from 'package:stats':  
##  
##   lowess
```

```

## Source the Random Forest wrapper script
source(paste0(root_directory, "Scripts/rfSubjectSpecific.R"))

## Set output directories for tables, figures and data structures.
if(!dir.exists(file.path(root_directory, "Figures"))){
  dir.create(file.path(root_directory, "Figures"))
}
if(!dir.exists(file.path(root_directory, "Tables"))){
  dir.create(file.path(root_directory, "Tables"))
}
if(!dir.exists(file.path(root_directory, "Script_output"))){
  dir.create(file.path(root_directory, "Script_output"))
}

R_script_output_directory<-paste0(root_directory, "Script_output/")
R_script_input_directory<-paste0(root_directory, "Script_input/") ## Contains
Rdata files created outside of this script. These data are static.
thesis_tables_directory<-paste0(root_directory, "Tables/")
thesis_figures_directory<-paste0(root_directory, "Figures/")

##Set global variables
seed_val<-4543 ## Needed for exact results obtained in thesis
pval_threshold<-0.05 ## Significance value threshold
npermutates<-500 ## Number of permutations to generate empirical null
distribution in RF models
nfolds<-10 ##Number of cross-fold validation in customized RF script
training_prop<-0.7 ## Proportion of input data for use in training models
nSpecies<-9 ## Max # of species to plot in Fig 1
sizes<-1 ## Default sizes for plots
raThreshold <-0.02 ## Min relative abundance threshold to plot in Fig 1
margins<-unit(c(-2.5,40,-2.5,5),units="points") ## Default margins for plots
ph_normalization_factor<-3 ## rescales y axis so that plot doesn't begin at 0
alpha_rect<-0.7 ## Fig 1 greyed out rectangle opacity
rect_fill<-"grey" ## Fig 1 greyed out color
removed_samples<-data.frame(Pre_QC_ID=NULL, QC_removal_stage=NULL)## Keep a
list of removed samples, and the stage of removal

## Set global plot theme
mBio <- theme_bw() + theme(text = element_text(family = "Arial", colour =
"black", size=12))

## Set standardized color table
load(file=paste0(R_script_input_directory, "subject_long_taxa_colors.Rdata"))
color_scheme_BCS<- c("L. crispatus"=
unnamed(subject_long_taxa_colors["Lactobacillus_crispatus"]),
           "L. jensenii" =
unnamed(subject_long_taxa_colors["Lactobacillus_jensenii"]),
           "L. iners" =
unnamed(subject_long_taxa_colors["Lactobacillus_iners"]),

```

```

        "G. vaginalis" =
unnamed(subject_long_taxa_colors["Gardnerella_vaginalis"]),
        "Cell Culture Medium"='blue'
) #c("L. crispatus"='red1',"G. vaginalis"='lightseagreen',"L.
iners"='orange',"Cell Culture Medium"='blue',"L. jensenii"='#8c510a')

cst.colors<-c("I-A"="red1",
              "I-B"="#990000",
              "I" =
unnamed(subject_long_taxa_colors["Lactobacillus_crispatus"]),
              "II"=unnamed(subject_long_taxa_colors["Lactobacillus_gasseri"]),
              "III-A"="darkorange" ,
              "III-B"="#cc7a00" ,
              "III"=unnamed(subject_long_taxa_colors["Lactobacillus_iners"]),
              "IV-A"="lightseagreen",
              "IV-B"="mediumblue",
              "V"=unnamed(subject_long_taxa_colors["Lactobacillus_jensenii"]),
              "DUMMY"='grey'
)

## Table and figure names
TABLE_SEQSUMMARY<-"TABLE_2_1.csv"
TABLE_MIR_TARGETS<-"TABLE_2_2.txt"
TABLE_TOPMIRS<-"TABLE_2_3.txt"
TABLE_QPCR_TIMECOURSE<-"TABLE_2_4.csv"
TABLE_EDU_SCRATCH_QUANT<-"TABLE_2_5.csv"
TABLE_CT_QUANT<-"TABLE_2_6.csv"

TABLE_TRL_ALIGNSTATS<-"TABLE_3_1.csv"
TABLE_TRL_NUMDEGENES<-"TABLE_3_2.csv"
TABLE_TRL_SUMMARY_PATHWAYS<-"TABLE_3_3.csv"

TABLE_COUNTS_RAW<-"TABLE_A4.csv"
TABLE_MODEL_INPUT<-"TABLE_A5.csv"
TABLE_SRL_METADATA<-"TABLE_A6.csv"
TABLE_PROXY_AMSEL_INPUT<-"TABLE_A7.csv"
TABLE_RF_SUMMARY.CV<-"TABLE_A8.csv"
TABLE_PROXY_AMSEL_SRL<-"TABLE_A9.csv"
TABLE_RF_SUMMARY<-"TABLE_A10.csv"

TABLE_TRL_COUNTS_RAW<-"TABLE_A11.csv"
TABLE_EDGER_RESULTS<-"TABLE_A12.csv"
TABLE_TRL_PATHWAY_Z_SCORES<-"TABLE_A13.csv"

FIGURE_SUBJECT_PLOTS<-"FIGURE_2_1_"
FIGURE_QC<-"FIGURE_2_2_"
FIGURE_QC_RIN_v_READS.PROP<-paste0(FIGURE_QC,"1.eps")
FIGURE_QC_RIN_v_READS.ABS<-paste0(FIGURE_QC,"2.eps")

```

```

FIGURE_QC_PCA.PREQC.BYBATCH<-paste0(FIGURE_QC,"3.eps")
FIGURE_QC_PCA.PREQC.BYSUBJ<-paste0(FIGURE_QC,"4.eps")
FIGURE_QC_PCA.RMLOW.BYBATCH<-paste0(FIGURE_QC,"5.eps")
FIGURE_QC_PCA.RMLOW.BYSUBJ<-paste0(FIGURE_QC,"6.eps")
FIGURE_QC_PCA.NORMAL.BYBATCH<-paste0(FIGURE_QC,"7.eps")
FIGURE_QC_PCA.NORMAL.BYSUBJ<-paste0(FIGURE_QC,"8.eps")
FIGURE_RF_IMPORTANCE_AMSEL<-"FIGURE_2_3_A.eps"
FIGURE_RF_IMPORTANCE_NUGENT<-"FIGURE_2_3_B.eps"
FIGURE_TOPMIRS<-"FIGURE_2_3_C.eps"
FIGURE_QPCR_TIMECOURSE<-"FIGURE_2_5.eps"
FIGURE_MIR_TARGETS_GO<-"FIGURE_2_4.eps"
FIGURE_DL_LACTICACID_QPCR<-"FIGURE_2_6.eps"
FIGURE_EDU_QUANT<-"FIGURE_2_7_D.eps"
FIGURE_SCRATCH_QUANT<-"FIGURE_2_7_C.eps"
FIGURE_CT_INFECT_QUANT<-"Figure_2_8_G.eps"
FIGURE_CT_EDU_QUANT<-"Figure_2_8_F.eps"
FIGURE_nyc_v_tsb<-"FIGURE_2_9.eps"

FIGURE_TRL_RIN_HIST<-"Figure_3_1.eps"
FIGURE_COMBINED_PATHWAYS_IMMUNE<-"FIGURE_3_2"
FIGURE_COMBINED_PATHWAYS_CYCLE<-"FIGURE_3_3"
FIGURE_LONGITUDINAL_GENEEXP.immune<-"FIGURE_3_4.eps"
FIGURE_LONGITUDINAL_GENEEXP.cycle<-"FIGURE_3_7.eps"

setwd(paste0(R_script_output_directory))

## SRL_meta and SRL_counts were prepared in separate script-
load(file=paste0(R_script_input_directory,"SRL_counts_meta.RData"))

## Extract counts data from ExpressionSet
SRL_counts_table<-exprs(SRL_counts_meta)
SRL_meta_table<-pData(SRL_counts_meta)

##Write SRL Raw Counts to disk
write.csv(SRL_counts_table,file=paste0(thesis_tables_directory,TABLE_COUNTS_R
AW),row.names=T,quote=F)

```

Custom functions

```

#####
## remove_poorQC_samples function
#####

## Adds samples in sample_list to running removed_samples data.frame. Tracks
the reason for removal
remove_poorQC_samples<-function(removed_samples=removed_samples,
                                sample_list=c(""),
                                reason=""){

removed_samples<-

```

```

unique(rbind(removed_samples,data.frame(Pre_QC_ID=sample_list,QC_removal_stag
e=reason)))
  return(removed_samples)
}

#####
## End remove_poorQC_samples function
#####

#####
## plot_pca function
#####

## Plots PCA and calculates gPCA-based p value. Outputs a list with PCA plot
object and p value (if plotly=F). Otherwise, just a plotly plot object

plot_pca<-function(ES, ## Expression Set object
  plot_title="", ## Title for PCA plot
  center=TRUE, ## pcrcomp centering?
  scale=TRUE, ##prcomp scaling?
  color_by='SID', ## What to color plots by
  logt=TRUE, ##Whether counts in ES should be Log transformed
  obs.scale = 1, ##prcomp scale factor
  var.scale = 1, ##prcomp var scaling
  ellipse = FALSE, ## draw ellipse around groups in PCA plot
  circle = FALSE, ## or circle
  var.axes=FALSE, ##option for prcomp
  plotly=FALSE, ##whether to generate a plotly interactive
plot. This option will not return gPCA p value
  margins=unit(c(0,0,0,0),units = "points"),
  seed_val=4543, ## seed needed for gPCA
  ...){

  ## es = ExpressionSet object
  ## ** Assumes ES counts data is log transformed. Set logt=FALSE to logt
data first.
  ## **Assumes columns named "Batch", "SID" ,"BVGroup" and "NUGENT_CLASS"
exist in pData in ES.

  set.seed(seed_val) ## Repeatable results in Guided PCA
  ##calculate variance of counts
  variable_counts<-t(exprs(ES))[,apply(t(exprs(ES)), 2, var, na.rm=TRUE) !=
0]

  ## log transform data if needed
  if(logt){
    cnts <- log(variable_counts+1,base = 2)
  }else{

```

```

    cnts<-variable_counts
  }

  ## extract metadata from ES to decorate tree. Assumes these variables are
  present
  batch <- as.numeric(pData(ES)[, 'Batch'])
  sid<-pData(ES)[, 'SID']
  bvgroup<-pData(ES)[, 'BVGroup']
  bvclass<-pData(ES)[, 'NUGENT_CLASS']
  ##color mappings for different metadata
  if(color_by=="SID"){
    group<-as.character(sid)
    group.gpca<-as.numeric(as.factor(group))
  }else if(color_by=="Batch"){
    group<-as.character(batch)
    group.gpca<-batch
  }else if (color_by=="NugentC"){
    group<-as.character(bvclass)
    group.gpca<-as.numeric(as.factor(group))
  }
  else{
    group<-as.character(bvgroup)
    group.gpca<-as.numeric(as.factor(group))
  }
  }

  ##PCA
  counts.pca <- prcomp(cnts,center = center, scale. = scale)

  #PCA plot
  pca.p <- ggbiplot(counts.pca, groups = group,
                   obs.scale = obs.scale, var.scale = var.scale ,
                   ellipse = ellipse,
                   circle = circle,var.axes =var.axes, ...)
  pca.p <- pca.p +
    scale_color_discrete(name = '')+
    mBio+
    theme(plot.margin = margins)

  gPCA.result<-gPCA.batchdetect(dist(cnts),group.gpca)

  ##Optional plotly functionality... returns a list with plot object in an
  element otherwise (with gPCA p value)

  if(plotly){ggplotly(pca.p)}else{return(list(pca.p=pca.p,gPCA.result=gPCA.resul
  t))}
  }

#####
## End plot_pca definition

```

```

#####

#####
## subset_ExpressionSet function
#####

## Function to subset ExpressionSet objects based on vector of sample/rows
names to remove. Returns an expression set without filterout samples

subset_ExpressionSet<-function(expSet, ##expression set
                               filterOut=c(""), ##vector of samples to drop
from expSet
                               samples=TRUE){ ## filter samples names or row
names

  ## Drop sample (columns) from expSet
  if(samples){
    ## Remove count data in filterOut vector
    counts_meta<-exprs(expSet)[,!colnames(exprs(expSet)) %in% filterOut]
    ## Remove metadata in filterOut vector
    design.subset<-pData(expSet)[!row.names(pData(expSet)) %in% filterOut,]
    ## Return a re-packaged filtered count and metadata into ExpressionSet
  }
  else{## Drop miRNAs (rows) from expSet
    counts_meta<-exprs(expSet)[!row.names(exprs(expSet)) %in% filterOut,]
    design.subset<-pData(expSet)[,!colnames(pData(expSet)) %in% filterOut]

  }
  return(ExpressionSet(assayData = as.matrix(counts_meta),
                      phenoData = AnnotatedDataFrame(design.subset)))
}

#####
## End subset_ExpressionSet
#####

#####
## plot_accuracy
#####

## Generates predicted vs actual tables and plots the predicted vs actual
values on plot

plot_accuracy<-function(rfp, ## Random Forest object
                       testing_fullset, ## vector of sample names that were
held out of training/used for testing

index_of_response=match("NUGENT_SCORE",names(testing_fullset)), ## index
corresponding to response variable in input data

```

```

        index_of_sid=match("SID",names(testing_fullset)),##
index corresponding to subject ID in input data (if subj_spec=T)
        subj_spec=TRUE, ## whether RF was run with
'subject_spec' option
        nfold=10) # number of k folds in RF if subj_spec=T
{

    accuracy_table<-data.frame(fold=0,predicted=0,actual=0) ## hold values for
actual, predicted values

    ## Subject specific rfp have a model for each cross fold, so need to loop
through and aggregate each fold
    if(subj_spec){

        for(m in 1:nfold){
            #m<-1
            ## Compare predicted to hold out set
            p1<-predict(rfp$mdl[[m]],
                testing_fullset[,-c(index_of_response,index_of_sid)],
type='response')

            accuracy_table<-
rbind(accuracy_table,data.frame(fold=m,predicted=p1,actual=testing_fullset[,i
ndex_of_response]))
        }
        }else{
            p1<-predict(rfp,testing_fullset[,-c(index_of_response,index_of_sid)],
type='response')
            accuracy_table<-
data.frame(fold="NA",predicted=p1,actual=testing_fullset[,index_of_response])
        }

        accuracy_table<-filter(accuracy_table,!fold==0) ## drop the initialization
row

        plot_a<-
ggplot()+geom_point(data=accuracy_table,aes(y=predicted,x=actual,col=as.facto
r(fold)))+
        ggtitle("Predicted vs Actual Values from plot_accuracy")+
        mBio+
        scale_x_continuous(limits=c(0,10),breaks = 1:10)+
        scale_y_continuous(limits=c(0,10),breaks = 1:10)
        print(plot_a)

        return(accuracy_table)
    }
}

```

```

#####
## End plot_accuracy definition
#####

#####
## run_randomForest definition
#####

## Wrapper to run Random Forest model building. Can run RandomForest,
rfPermute or rfSubjectSpecific. Takes care of subsetting data for
training/testing and outputting accuracy/error/etc estimates. Saves/loads
previous models.
#Note if load_prev=T, most of the parameters are ignored
run_randomForest<-function(predictors_response_table,
                             response_variable_name,
                             nfold=10,
                             nreps=105,
                             permute=TRUE,
                             save_model=FALSE,
                             load_prev_model=TRUE,
                             file_n="rf_model",
                             verbose=TRUE,
                             pval_thres=pval_threshold,
                             subj_spec=TRUE,
                             importance_thres=10,
                             training_prop=0.7,
                             seed=seed_val,
                             R_script_output_directory,
                             ...)
{
  # predictors_response_table - data frame of predictors + response variable.
If it contains a column called 'SID', sets subj_spec=TRUE
  # response_variable_name - column name as string of response variable
(as found in predictors_response_table)
  # nfold - number of k-fold validations to run
  # nreps - number of permutations to run to compute null
distribution. Ignored if permute=FALSE
  # permute - whether or not to run rfPermute (generate
null distribution permutation p-values for each feature)
  # save_model - save model as Rdata to outout directory? Uses
'file_n' as file name
  # load_prev_model - Load model outout directory? Uses 'file_n' as
file name
  # file_n - file name to use when reading or writing a
model to disk
  # verbose - detailed output of model results, etc
  # pval_thres -p value threshold to call signifigant
  # subj_spec - whether or not to run rfSubjectSpecific.R.
Looks for a column called 'SID' in predictors_response_table

```

```

# importance_thres           -
#training_prop              - proportion of input data to use as training.
remaining is used as hold out set for testing
# seed                      -seed value to produce repeatable results
# R_script_output_directory - root output directory for read/write model
files

nrep<-nreps
set.seed(seed)
## First major control point: loading from previously saved model, or
generating a new trained model?

#### //////////
# Start New Model
#### //////////

if(!load_prev_model){

  #Find the column index corresponding to response & subject ID
  index_of_response<-
match(response_variable_name,names(predictors_response_table))
  index_of_sid<-match("SID",names(predictors_response_table))
  if(is.na(index_of_sid) & subj_spec){
    print("SID could not be found. Setting to non-subject specific")
    subj_spec<-FALSE
    index_of_sid<-0
  }else if (is.na(index_of_sid)){
    index_of_sid<-0
  } else if(index_of_sid>0 & !subj_spec){
    print("Found a SID column. Setting to subject-specific. ")
    subj_spec<-TRUE
  }

  #Partition input data into training and testing

  inTrain<-
createDataPartition(y=predictors_response_table[,index_of_response],p
=train_prop,list = F)
  if(is.character(predictors_response_table[,index_of_response])){
    predictors_response_table[,index_of_response]<-
as.factor(predictors_response_table[,index_of_response])
  }

  ##Subjet training and testing data
  training_fullset<-predictors_response_table[inTrain,]
  testing_fullset<-predictors_response_table[-inTrain,]
  table(predictors_response_table[,index_of_response])
  table(training_fullset[,index_of_response])
  table(testing_fullset[,index_of_response])

```

```

response<-training_fullset[,index_of_response]

# Determine whether to run classification or regression depending on
response variable type
if(is.numeric(predictors_response_table[,index_of_response])){#
regression
  rf_type<-"regression"
}else{
  rf_type<-"classification"
}

#Determine whether to run rfSubjectSpecific.R
if(subj_spec){
  print(paste0("Starting subject-specific rfSubjectSpecific with permute
set to : ",permute))
  rfp<-rfSubjectSpecific(training_fullset[, -
c(index_of_response,index_of_sid)],response,subjID =
as.character(training_fullset[,index_of_sid],nrep=nrep),nfolds =
nfold,verbose=verbose,nrep = nreps,permute=permute) ## This will be sourced
at the setup section. It is an external script.
}else{
  print("Starting non subject-specific rfPermute")
  rfp<-rfPermute(training_fullset[, -c(index_of_response)],response,nrep =
nreps,...)
}

## The accuracy, etc output varies depending on whether which combination
of RF were run

## /////
## Regression + Subject-Specific
## ////

if(rf_type=="regression" & subj_spec){ ## If it's a regression model
  accuracy_table<-plot_accuracy(rfp,testing_fullset =
testing_fullset,index_of_response = index_of_response, index_of_sid =
index_of_sid,subj_spec = subj_spec,nfold = nfold) ## see above for this
function

  ## /////
  ## Classification + Non Subject-Specific
  ## ////

}else if (rf_type=="classification" & !subj_spec){
  p1<-predict(rfp, testing_fullset[, -c(index_of_response,index_of_sid)],
subj_spec = subj_spec,type='response')
  (accuracy_table<-
table(Var1=p1,Var2=testing_fullset[,index_of_response]))

```

```

## /////
## Classification + Subject-Specific
## ////

}else if (rf_type=="classification" & subj_spec){
  p1<-predict(rfp$mdl, testing_fullset[, -
c(index_of_response,index_of_sid)], subj_spec = subj_spec,type='response')
  (accuracy_table<-lapply(p1, function(x)
table(x,testing_fullset[,c(index_of_response)])))

  ## /////
  ## All else (Regression + Non Subject-Specific)
  ## ////

}else{
  accuracy_table<-NULL
}

## Write model to file + accuracy table and training/testing info

if(save_model){
  save(rfp,file=paste0(R_script_output_directory,file_n,".RData"))

save(accuracy_table,file=paste0(R_script_output_directory,file_n,"_accuracyTable.RData"))
  training_testing<-
list(training_fullset=training_fullset,testing_fullset=testing_fullset)

save(training_testing,file=paste0(R_script_output_directory,file_n,"_training_testing.RData"))
}

#### //////////
# Load Previous Model
#### //////////
}else{

  rfp.pointer<-load(file=paste0(R_script_output_directory,file_n,".RData"))
  rfp<-get(rfp.pointer)

  accuracy_table.pointer<-
load(file=paste0(R_script_output_directory,file_n,"_accuracyTable.RData"))
  accuracy_table<-get(accuracy_table.pointer)

  training_testing.pointer<-
load(file=paste0(R_script_output_directory,file_n,"_training_testing.RData"))
  training_testing<-get(training_testing.pointer)
  training_fullset<-training_testing$training_fullset
  testing_fullset<-training_testing$testing_fullset

```

```

}
if(verbose){
  print(rfp)
  print(rfp$mdl)
}

if(permute & subj_spec){
  importance<-data.frame(rfp$importance) ##only happens if
rfSubjectSpecific is run with permute
  importance.pval<-dplyr::select(importance,ends_with("pval"))
}else{
  if(!subj_spec){
    importance<-data.frame(rp.importance(rfp))
  }else{
    importance<-data.frame(rfp$imp)
  }
}
return(list(rfp=rfp, ##model
           accuracy_table=accuracy_table,
           importance=importance,
           training_ids=row.names(training_fullset),
           testing_ids=row.names(testing_fullset)))
}

#####
## END run_randomForest function
#####

#####
## plot_RIN
#####

## Is there an effect due to RIN and number of reads miRNAs?
### # Wrapper for plot of reads miRNAs vs RIN

plot_RIN_meta<-function(SRL_meta_table, ## Table containing SRL metadata
                        y_series="reads_surviving.percent", ## y axis column
                        name to be plotted
                        col_by="Batch", ## color by
                        vjust = 0, ## vertical justification
                        nudge_y = 0.05, ## nudge y by...
                        angle = 0, ## y axis angle
                        hjust = 0, ## hortizontal adjustment
                        nudge_x = 0.05, ## nudge x by..
                        check_overlap = FALSE, ## check for points overlap-
try to minimize
                        y_series_label=y_series # y Label on plot
){

```

```

## Plot for RIN vs a measure of read mapping
p.RIN<-
ggplot(SRL_meta_table,aes(x=RIN,y=SRL_meta_table[,y_series],col=as.factor(SRL
_meta_table[,col_by]),label=SRL_meta_table$Pre_QC_ID),label = Pre_QC_ID)+
  geom_point()+
  ylab(y_series_label)+
  xlab("RINe")+
  # ggtitle(paste(y_series, " vs RIN"))+
  geom_text(check_overlap = check_overlap,vjust = vjust, nudge_y =
nudge_y,angle = angle,hjust = hjust, nudge_x = nudge_x)+
  mBio+
  guides(col=guide_legend(title=col_by))
return(p.RIN)
}

#####
## end plot_RIN
#####

#####
## mapping_stats
#####

## Generate simple mapping stats given an Expression Set pData object
mapping_stats<-function(column=pData(counts_meta.qc)$number_reads_mirs){
  return(data.frame(min=min(column,na.rm = T),median=median(column,na.rm =
T),max=max(column,na.rm = T)))
}

#####
## end mapping_stats
#####

#####
## plot_importance
#####

## Plot RF Importance variables. Output is a plot

plot_importance<-function(importance_df=Nugent_RF$importance, ## Data frame
holding importance results

ntopfeats=length(Nugent_RF$top_features$top_features.all), ## number of
features to plot
  nfeats=25,#max # of features to plot
  rankBy="IncMSE", ## Importance variable to rank by
  model_name="",
  size_font=12,

```

```

                                size_points=5
){
  if(nfeats<ntopfeats){
    stop("Number of top features less than total # of features plotted")
  }
  names(importance_df)<-gsub(pattern = "X\\.",replacement = "",x =
names(importance_df))
  importance_df<-importance_df[order(importance_df[,rankBy],decreasing = T),]
  importance_df.cut<-importance_df[1:nfeats,]
  importance_df.cut$features<-factor(row.names(importance_df.cut),levels =
rev(row.names(importance_df.cut)),ordered = T)

  importance_df.cut.tmp<-separate(melt(importance_df.cut,id.vars =
c("features")),col = variable,into=c("metric","SUFFIX"),sep = "\\.",fill =
"right")

  importance_df.cut.tmp[is.na(importance_df.cut.tmp$SUFFIX),"SUFFIX"]<-
"metric_val"
  importance_df.cut<-spread(data=importance_df.cut.tmp,key = SUFFIX,value =
value)
  cutoff<-nfeats+0.5-ntopfeats

  ggplot(importance_df.cut)+
    geom_point(aes(y=features,x=metric_val,col=-
log(pval,10),pch=metric),size=size_points)+
    xlab("Importance Metric")+
    ylab("Predictor")+
    scale_colour_gradient(high="red", low="blue",guide = guide_legend(title =
"-Log(p-value)"))+
    mBio+
    theme(text=element_text(size = size_font),panel.border =
element_rect(size=1))+
    ggtitle(paste(model_name," RF Importance Plot for the
First",nfeats,"Ranked Features (by",rankBy,")"))
}

#####
## end plot_importance
#####

#####
## t.test2
#####

## A t test using sumamry statistics ( and not the entire sample dataset as
with t.test)
## Returns p value, difference of means, standard error and t statistc
t.test2 <- function(m1, ## mean of sample set 1

```

```

        m2, ## mean of sample set 2
        s1, ## standard dev of sample set 1
        s2, ## standard dev of sample set 2
        n1, ## number of samples in sample set 1
        n2, ## number of samples in sample set 1
        m0=0, ## the null for hypothesis to test (mean value)=
        equal.variance=FALSE) #whether to assume equal variance
between sample sets

{
  if( equal.variance==FALSE )
  {
    ## "normalize" standard deviations if unequal variance to compute
    standard error
    se <- sqrt( (s1^2/n1) + (s2^2/n2) )
    # welch-satterthwaite df
    df <- ( (s1^2/n1 + s2^2/n2)^2 )/( (s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1)
  )
  } else
  {
    # pooled standard deviation, scaled by the sample sizes
    se <- sqrt( (1/n1 + 1/n2) * ((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2) )

    df <- n1+n2-2
  }
  t <- (m1-m2-m0)/se## calculate t statistic
  dat <- c(m1-m2, se, t, 2*pt(-abs(t),df)) ## calculate p value based on
Student's t distribution
  names(dat) <- c("Difference of means", "Std Error", "t", "p-value")
  return(dat)
}

#####
## end t.test2
#####

#####
## apply_ttest
#####

apply_ttest<-
function(sigtest_results=proliferation_sigtest,pval_threshold=pval_threshold,
summary_stats,test_function="t.test",value="value",obs=""){

  ## Loop through each test comparison and apply stat test. Store in
  dataframe
  for(sigtest in 1:nrow(sigtest_results)){
    print(sigtest)

```

```

#sigtest<-1
x1<-as.character(sigtest_results[sigtest,"xref"])
y1<-as.character(sigtest_results[sigtest,"reference"])

## Apply a t test to x1 vs y1
if(test_function=="t.test"){
  tes<-t.test(x=filter(summary_stats,BCS==x1
)$percent_cells,y=filter(summary_stats,BCS==y1)$percent_cells,alternative =
"two.sided")
  delta_mean<-tes$estimate[2]-tes$estimate[1]
  pval<-tes$p.value
}else if(test_function=="t.test2"){
  obs<-as.character(sigtest_results[sigtest,"Observation"])
  m1<-filter(summary_stats,Observation==obs & Treatment==x1 )$grand_mean
  m2<-filter(summary_stats,Observation==obs & Treatment==y1 )$grand_mean
  s1<-filter(summary_stats,Observation==obs & Treatment==x1 )$grand_sd
  s2<-filter(summary_stats,Observation==obs & Treatment==y1 )$grand_sd
  n1<-filter(summary_stats,Observation==obs & Treatment==x1 )$n
  n2<-filter(summary_stats,Observation==obs & Treatment==y1 )$n
  print(obs)
  tes<-t.test2(m1 = m1,s1 = s1, m2=m2,s2=s2,n1=n1,n2=n2) #Less
  print(tes)
  delta_mean<-tes["Difference of means"]
  pval<-tes["p-value"]

}else{
  print(paste0(test_function," not found"))
}

## Store p value and difference of means
sigtest_results[sigtest,"pval"]<-pval
sigtest_results[sigtest,"mean_diff"]<-delta_mean
}

## Denote significant tests with "*"
sigtest_results[sigtest_results$pval<=pval_threshold,"sig"]<-"*"
return(sigtest_results)
}

#####
## end apply ttest
#####

#####
## setup_sigtest
#####
#pval_threshold = pval_threshold;raw_data = proliferation.m;test_function =
"t.test";Experiment = "Scratch"

```

```

setup_sigtest<-function(pval_threshold = pval_threshold,raw_data =
raw_data,test_function = "t.test",Experiment="Scratch"){
  ## Set up a significance test df to store inference testing data
  if(Experiment=="Scratch"){
    sigtest<-data.frame(xref=c(rep("L. crispatus",times=4),rep("L.
jensenii",times=3),rep("L. iners",times=2),"G. vaginalis"),reference=c("L.
jensenii","L. iners","G. vaginalis","Cell Culture Medium","L. iners","G.
vaginalis","Cell Culture Medium","G. vaginalis","Cell Culture Medium","Cell
Culture Medium"))

    ## Factor labels
    sigtest$xref<-factor(sigtest$xref,levels =c("L. crispatus","L.
jensenii","L. iners","G. vaginalis","Cell Culture Medium"),ordered = T)
    sigtest$reference<-factor(sigtest$reference,levels = c("L. jensenii","L.
iners","G. vaginalis","Cell Culture Medium"),ordered = T)

    sigtest<-apply_ttest(sigtest_results = sigtest,pval_threshold =
pval_threshold,summary_stats = raw_data,test_function = "t.test")
    ## Summarize scratch assay data with mean and sd
    summary_stats<-
ddply(dplyr::select(raw_data,c(BCS,percent_cells)),c("BCS"),summarise,mean=me
an(percent_cells),sd=sd(percent_cells))

  }else if(Experiment=="Infection"){

    sigtest<-
data.frame(Observation=rep(c("Proliferation","Infectivity"),each=3),xref=c(re
p("CAS 546102-60-
7",times=2),rep("Fascaplysin",times=1)),reference=c("Fascaplysin","Cell
Culture Medium","Cell Culture Medium"))

    sigtest$xref<-factor(sigtest$xref,ordered = T,levels = c("CAS 546102-60-
7","Fascaplysin"))
    sigtest$reference<-factor(sigtest$reference,ordered = T,levels =
c("Fascaplysin","Cell Culture Medium"))

    raw_data.spread<-spread(raw_data,key = Observation,value = percent_cells)

plot(raw_data.spread$Infectivity,raw_data.spread$Proliferation,col=raw_data.s
pread$Treatment)

    summary_stats.pre<-
ddply(raw_data,c("Observation","Treatment","Coverslip"),summarise,mean=mean(p
ercent_cells,na.rm = T),sd=sd(percent_cells,na.rm = T))

    summary_stats<-
ddply(summary_stats.pre,c("Treatment","Observation"),summarise,grand_mean=mea

```

```

n(mean,na.rm = T))

  for(obs in c("Infectivity","Proliferation")){
    for(treat in unique(summary_stats$Treatment)){
      #treat<-"Control"
      gm<-summary_stats[summary_stats$Observation==obs &
summary_stats$Treatment==treat,"grand_mean"]
      means<-summary_stats.pre[summary_stats.pre$Treatment==treat &
summary_stats.pre$Observation==obs,"mean"]
      n<- length(means[!is.na(means)])
      summary_stats[summary_stats$Treatment==treat &
summary_stats$Observation==obs,"grand_sd"]<-sqrt(sum((means-gm)^2,na.rm =
T)/(n-1))
      summary_stats[summary_stats$Observation==obs &
summary_stats$Treatment==treat,"n"]<-n
    }
  }

  sigtest<-apply_ttest(sigtest_results = sigtest,pval_threshold =
pval_threshold,summary_stats = summary_stats,test_function = "t.test2")

}else{
  print("Experiment must be Infection or Scratch")
}

statbars<-data.frame(xref=sigtest$xref,reference=sigtest$reference,
  x=as.numeric(sigtest$xref)+.05,
  xend=(as.numeric(sigtest$reference)+1)-.05)
if(Experiment=="Scratch"){

  statbars[statbars$reference=="Cell Culture Medium","y"]<-rep(seq(100,91,-
3),times=1)
  statbars[statbars$reference=="G. vaginalis","y"]<-rep(seq(91,by=-
3,length.out = 3),times=1)
  statbars[statbars$reference=="L. iners","y"]<-rep(c(85,94),times=1)
  statbars[statbars$reference=="L. jensenii","y"]<-rep(94,times=1)
  statbars[statbars$reference=="Cell Culture Medium" & statbars$xref=="L.
iners","y"]<-85
  statbars[statbars$reference=="G. vaginalis" & statbars$xref=="L.
iners","y"]<-94

}else if(Experiment=="Infection"){

  statbars[statbars$reference=="Cell Culture Medium","y"]<-rep(seq(100,97,-
3),times=2)
  statbars[statbars$reference=="Fascaplysin","y"]<-rep(c(91,94),times=1)
  statbars[statbars$reference=="CAS 546102-60-7","y"]<-rep(90,times=2)

```

```

statbars[statbars$xref=="CAS 546102-60-7"
&statbars$reference=="Fascaplysin" ,"y"] <-97

}else{
  print("Experiment must be Infection or Scratch")
}

statbars$yend<-statbars$y
statbars2<-data.frame(x=c(statbars$x,statbars$xend),
                      y=rep(statbars$y,times=2),
                      yend=rep(statbars$y-1.5,times=1))
statbars2$xend<-statbars2$x

sigtest<-merge(statbars,sigtest,by = c("xref","reference"))

sigtest$midpoints<-((sigtest$xend-sigtest$x)/2)+sigtest$x
sigtest$y.sig<-sigtest$y+1
sigtest$sig[is.na(sigtest$sig)]<-"N.S."

return(list(sigtest=sigtest,statbars=statbars,statbars2=statbars2,summary_stats=summary_stats))
}

#####
## end setup_sigtest
#####

#####
## plot_replicates
#####

plot_replicates<-
function(eset,BCS.selection=c(""),ExposureTime.selection=c(""),logt=F,rmlow=F
,lowcnt=10){

  if(logt){
    cnts<-log(exprs(eset)+1,base = 2)
  }else{
    cnts<-exprs(eset)
  }
  if(rmlow){

    cnts<-cnts[rowSums(cnts>lowcnt)==ncol(cnts),]
  }

  pairs.panels(cnts[, pData(eset)$BCS %in% c(BCS.selection) &

```

```
pData(eset)$ExposureTime %in% c(ExposureTime.selection)]
}
```

```
#####
## end plot_replicates
#####
```

```
#####
## subset_ExpressionSet
#####
```

Summary Statistics & Quality Control

Alignment Stats

Summarize the SmallRNA seq alignments

```
## Stats on # reads hg (% surviving), # reads miRNA (% surviving, hg)
SRL_meta_table_summary_table<-
data.frame(matrix(0,nrow=3,ncol=3));names(SRL_meta_table_summary_table)<-
c("min","median","max");row.names(SRL_meta_table_summary_table)<-
c("trimmed","hg19","miRNome")

SRL_meta_table_summary_table["trimmed",]<-
summarise(SRL_meta_table,min=min(reads_surviving),median=median(reads_survivi
ng),max=max(reads_surviving))
SRL_meta_table_summary_table["hg19",]<-
summarise(SRL_meta_table,min=min(hg19.mapped),median=median(hg19.mapped),max=
max(hg19.mapped))
SRL_meta_table_summary_table["miRNome",]<-
summarise(SRL_meta_table,min=min(number_reads_mirs),median=median(number_read
s_mirs),max=max(number_reads_mirs))
print(SRL_meta_table_summary_table)

##           min   median      max
## trimmed 3225 27087178 107477660
## hg19     607  4499861  17195821
## miRNome  159  1102119   4642910
```

Quality Control (QC) small RNA-seq count data

```
print(paste("Number of samples in initial counts
table:",ncol(SRL_counts_table)))

## [1] "Number of samples in initial counts table: 113"

print(paste("Number of miRNAs in initial counts
table:",nrow(SRL_counts_table)))
```

```

## [1] "Number of miRNAs in initial counts table: 1869"

## Determine number of miRNAs without any counts across all samples
missing<-rowSums(SRL_counts_table)==0
#rowSums(is.na(SRL_counts_table))==ncol(SRL_counts_table) |
print(paste("Number of miRNAs without any counts across
all",ncol(SRL_counts_table),"samples:",sum(missing)))

## [1] "Number of miRNAs without any counts across all 113 samples: 323"

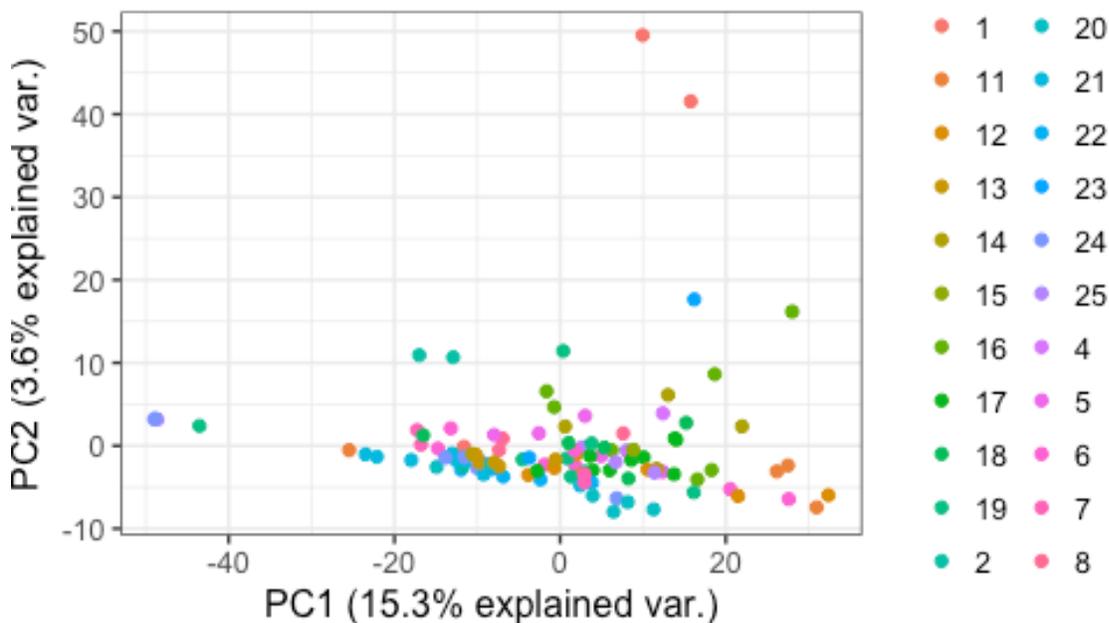
print(paste0("which leaves ",100*(1-sum(missing)/nrow(SRL_counts_table)),"
percent of all annotated miRNAs with at least 1 read"))

## [1] "which leaves 82.7180310326378 percent of all annotated miRNAs with at
least 1 read"

#print("...and the names of the miRNAs without any reads across all samples:
")
#names(missing)[missing]

plot_QC.preQC.batch<-plot_pca(SRL_counts_meta,"Log miRNA Raw Count Distance,
PRE QC",color_by = "Batch",plogy = F,seed_val = seed_val)##FIGURE
plot(plot_QC.preQC.batch$pca.p)

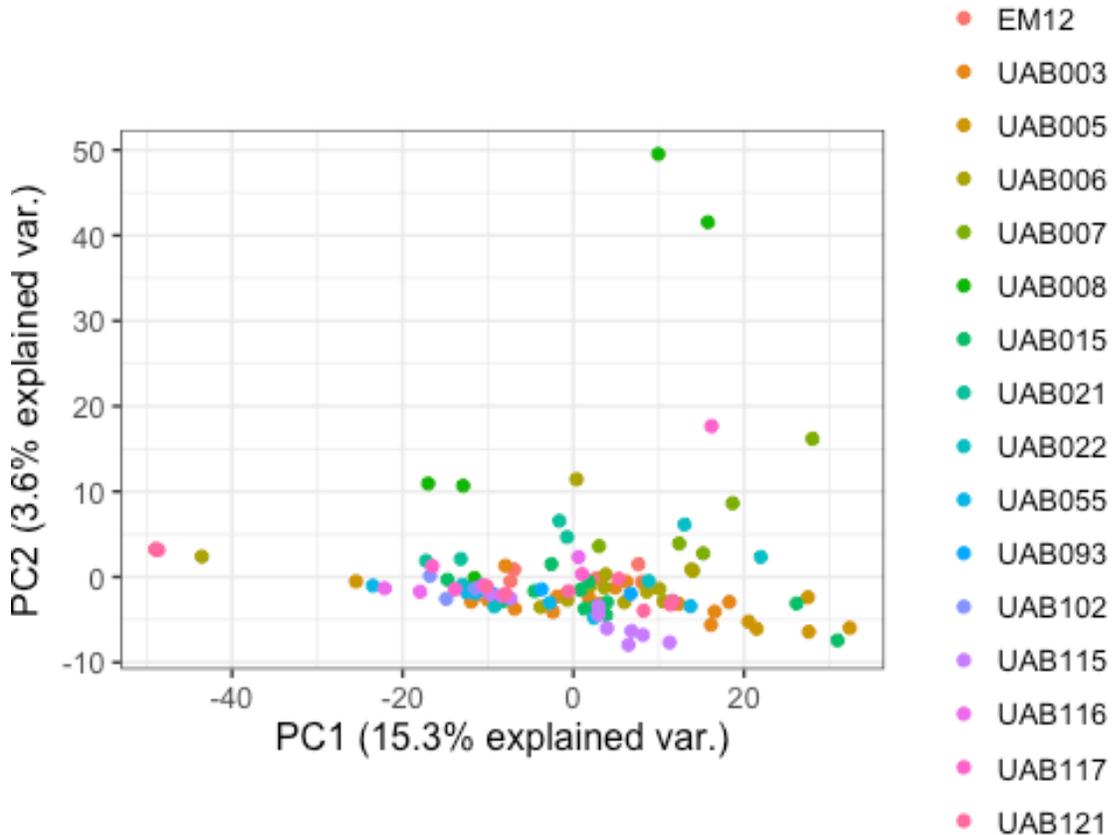
```



```
plot_QC.preQC.batch$gPCA.result$p.val
```

```
## [1] 0.994
```

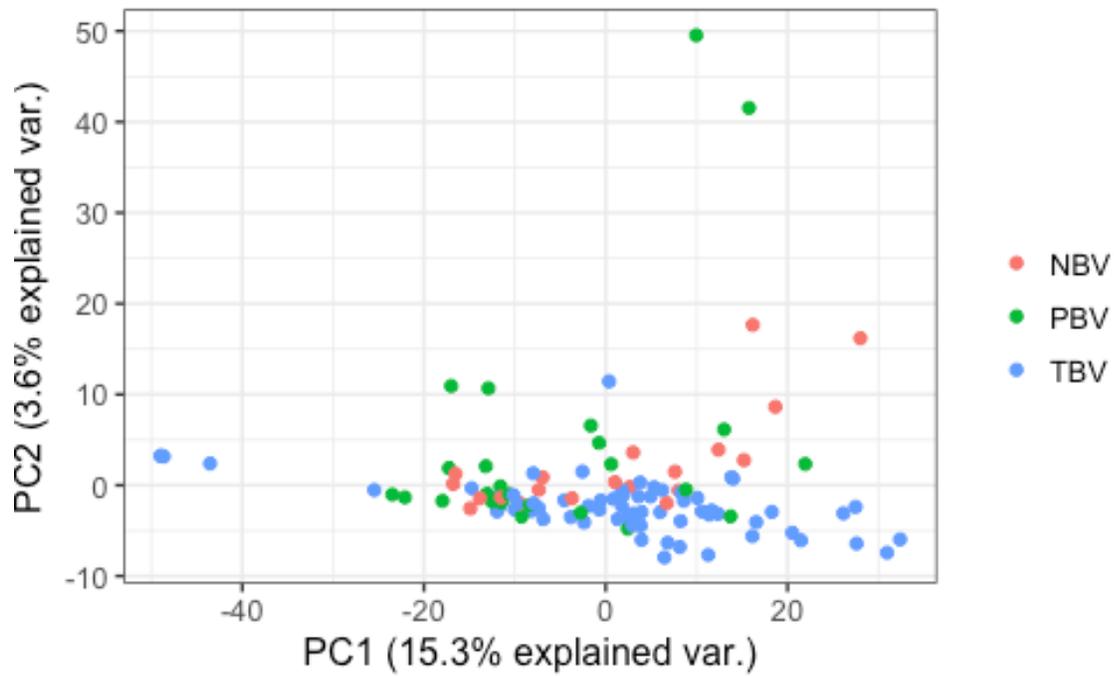
```
plot_QC.preQC.sid<-plot_pca(SRL_counts_meta,"Log miRNA Raw Count Distance,  
PRE QC",color_by = "SID",plogy = F,seed_val = seed_val)##FIGURE  
plot(plot_QC.preQC.sid$pca.p)
```



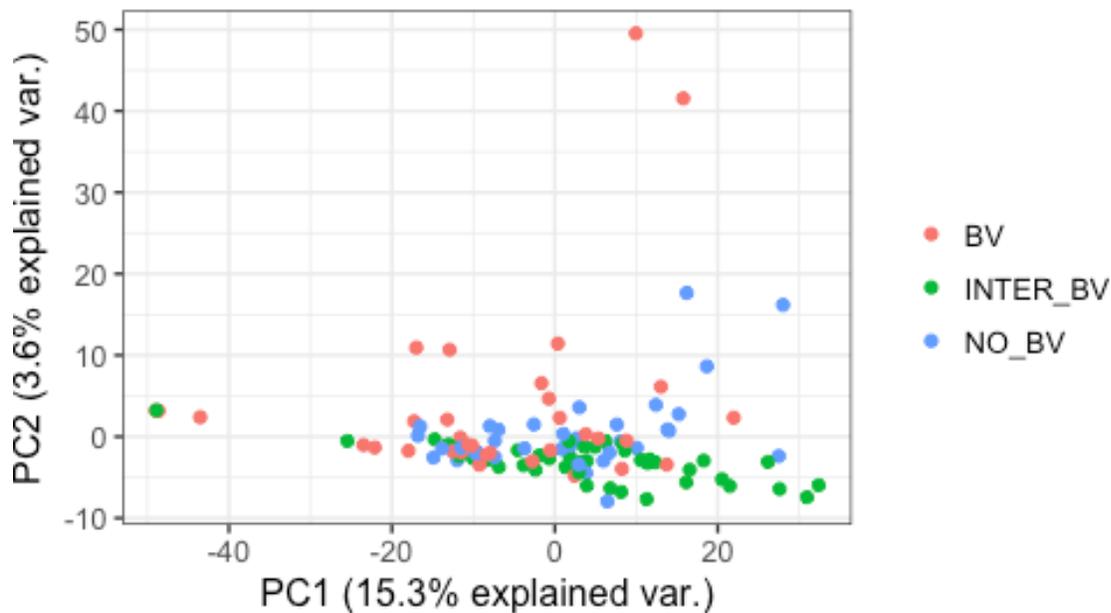
```
plot_QC.preQC.sid$gPCA.result$p.val
```

```
## [1] 0.288
```

```
plot_pca(SRL_counts_meta,"Log miRNA Raw Count Distance, PRE QC",color_by =  
"BVGroup",seed_val = seed_val)$pca.p##FIGURE
```



```
plot_pca(SRL_counts_meta,"Log miRNA Raw Count Distance, PRE QC",color_by =
"NugentC",seed_val = seed_val)$pca.p##FIGURE
```



```
## Get a *ROUGH* idea of miRNA coverage by dividing counts by # of annotated
miRNAs in genome
(number_mirs_genome<-nrow(SRL_counts_meta)) ## number of annotated miRNAs

## Features
##      1869

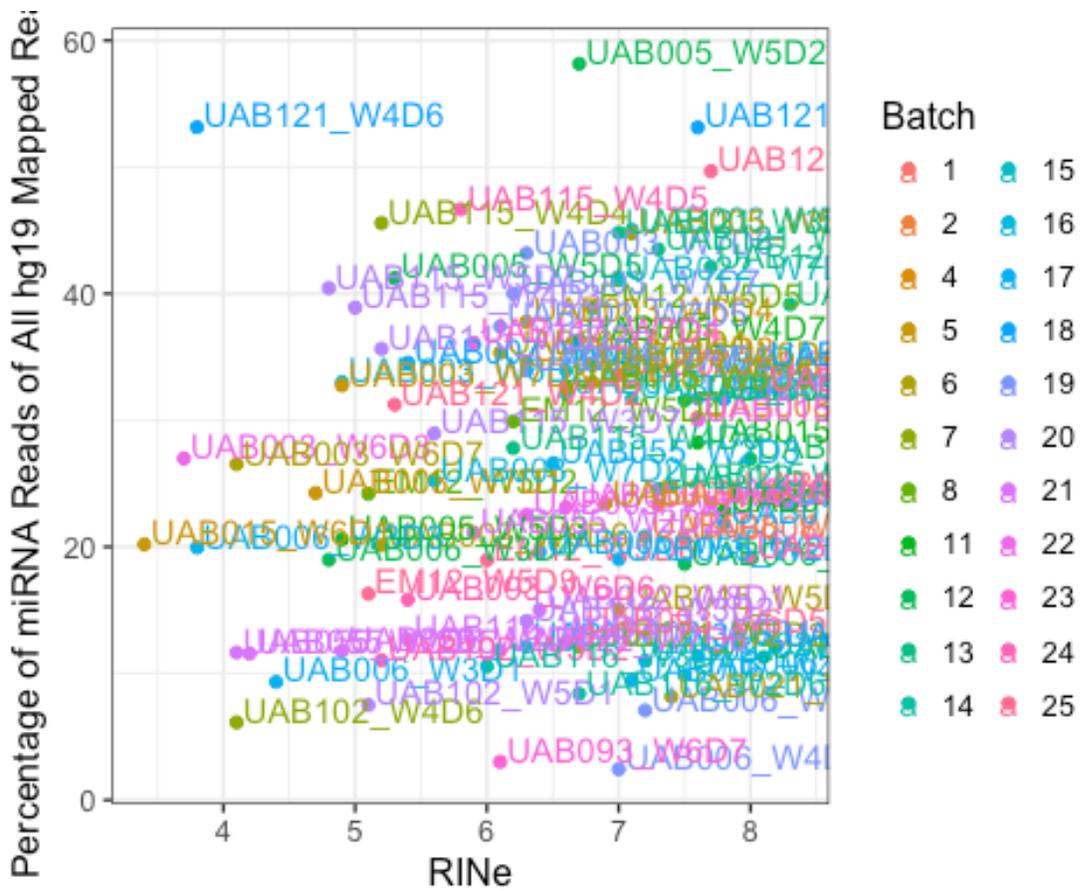
#SRL_counts_table<-SRL_counts_table[!missing,] ## non missing counts only
SRL_counts_meta<-SRL_counts_meta[!missing]
sum(missing)

## [1] 323

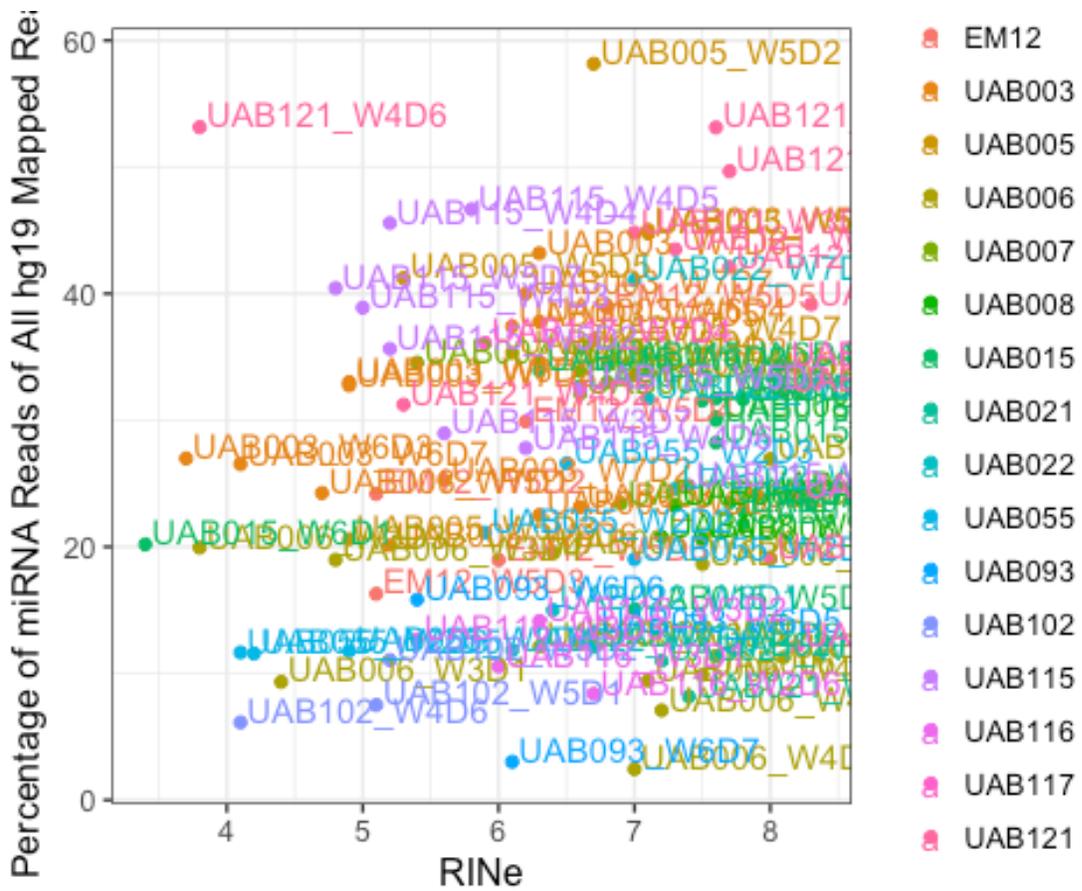
## Number of miRNAs completely removed due to non-detection
nrow(SRL_counts_meta)

## Features
##      1546

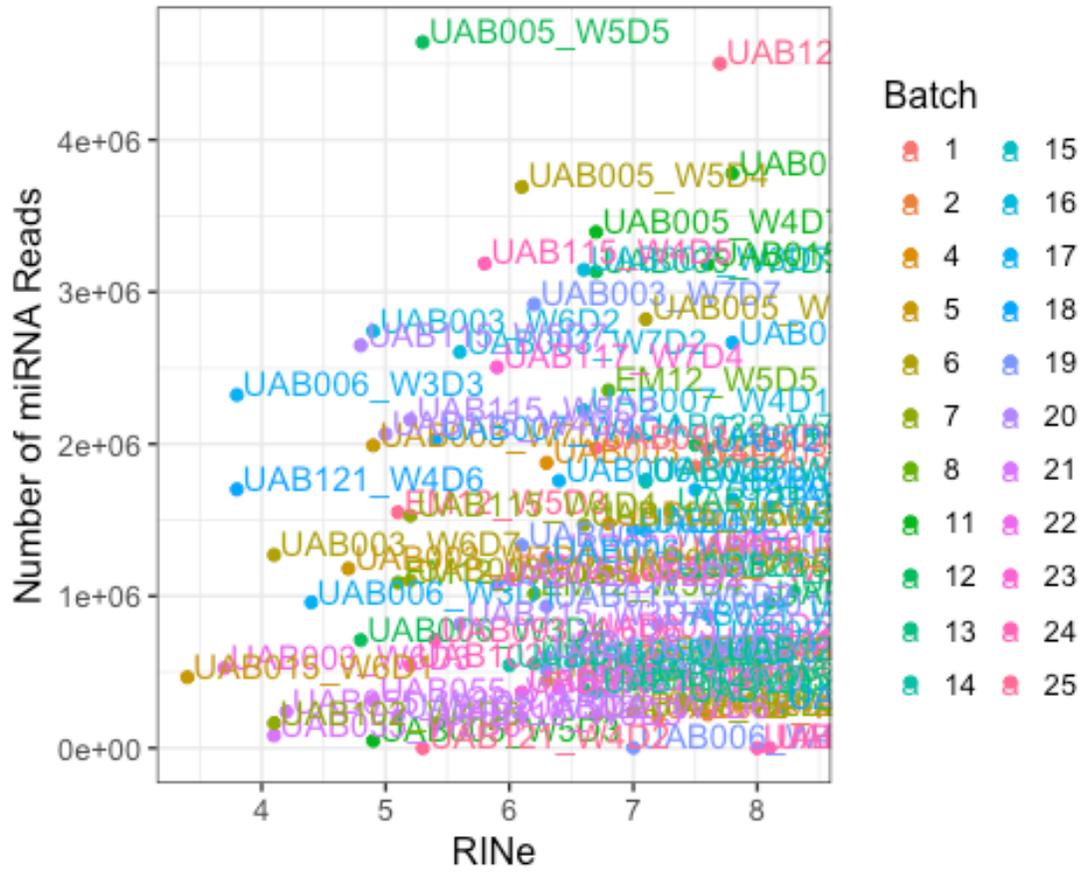
(plot_QC.RIN.percent.batch<-
plot_RIN_meta(SRL_meta_table,"number_reads_mirs.percent_hgmapped",col_by="Batch",y_series_label = "Percentage of miRNA Reads of All hg19 Mapped Reads"))##FIGURE
```



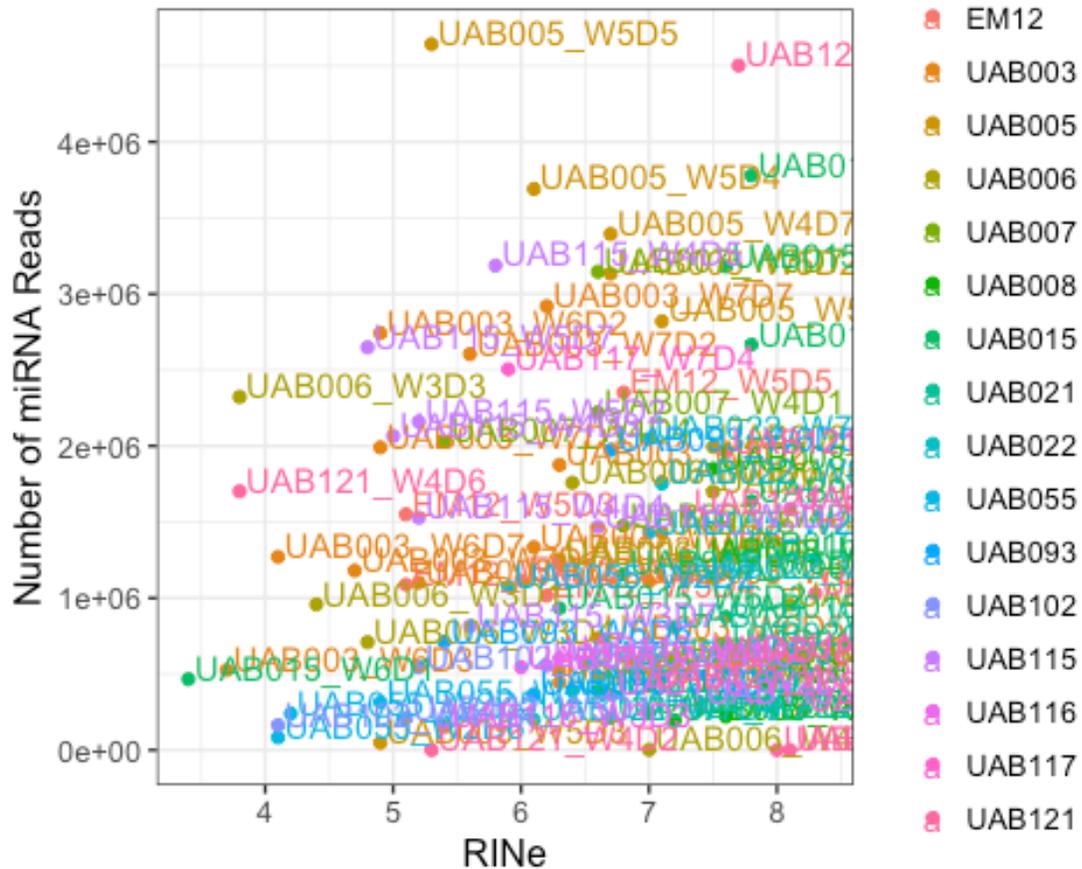
```
(plot_QC.RIN.percent.sid<-
plot_RIN_meta(SRL_meta_table,"number_reads_mirs.percent_hgmapped",col_by="SID
",y_series_label = "Percentage of miRNA Reads of All hg19 Mapped Reads"))
```



```
(plot_QC.RIN.number.batch<-
plot_RIN_meta(SRL_meta_table,"number_reads_mirs",col_by="Batch",y_series_label = "Number of miRNA Reads"))##FIGURE
```



```
(plot_QC.RIN.number.sid<-
plot_RIN_meta(SRL_meta_table,"number_reads_mirs",col_by="SID",y_series_label
= "Number of miRNA Reads"))
```



```
## Range of miRNA coverage based on simple reads/#annotated miRNAs
print(summary(colSums(SRL_counts_table))/number_mirs_genome)
```

```
##      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
##  0.0851  258.0524  589.6201  684.3232  989.8341 2484.2162
```

```
## Average coverage
```

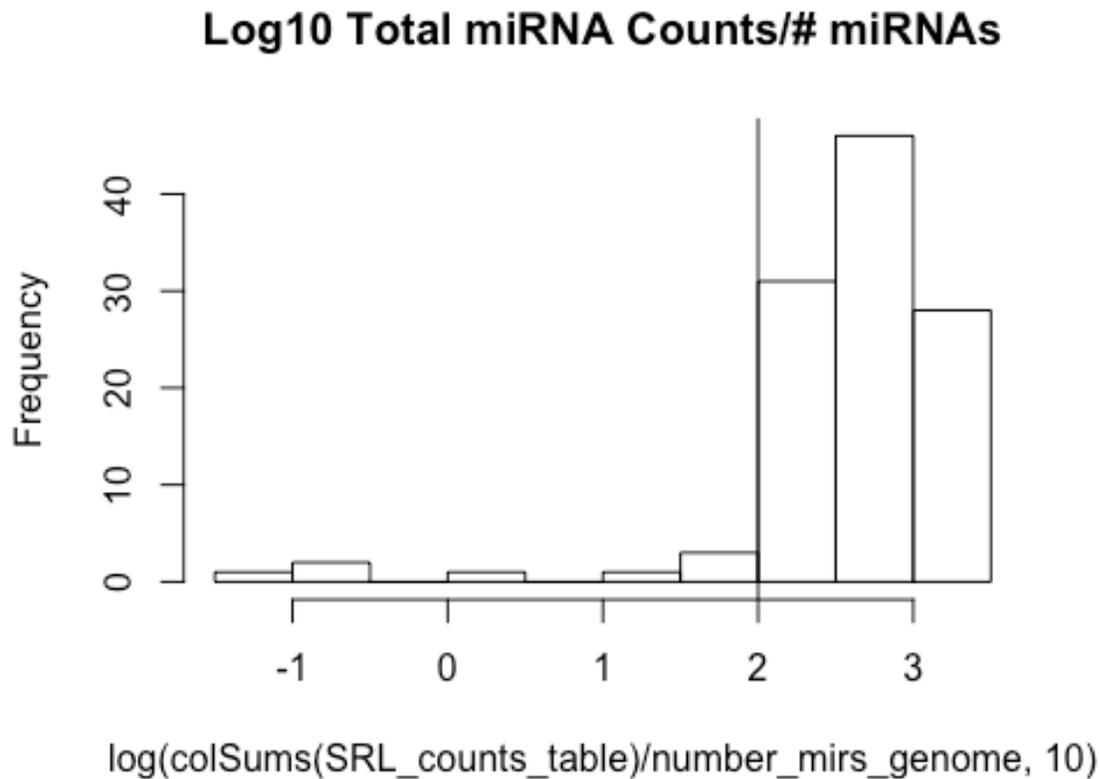
```
data.frame(Pre_QC_ID=SRL_meta_table$Pre_QC_ID,avg_coverage=SRL_meta_table$number_reads_mirs/number_mirs_genome)
```

```
##      Pre_QC_ID avg_coverage
## 1      EM12_W5D2 5.821814e+02
## 2      EM12_W5D3 8.298529e+02
## 3      EM12_W5D4 5.437319e+02
## 4      EM12_W5D5 1.258753e+03
## 5      EM12_W5D6 5.969540e+02
## 6      UAB003_W5D6 9.440787e+02
## 7      UAB003_W6D1 6.720284e+02
## 8      UAB003_W6D2 1.467211e+03
## 9      UAB003_W6D3 2.817180e+02
## 10     UAB003_W6D4 2.377742e+02
## 11     UAB003_W6D5 3.938844e+02
## 12     UAB003_W6D6 5.896838e+02
## 13     UAB003_W6D7 6.799176e+02
```

14 UAB003_W7D1 6.317940e+02
15 UAB003_W7D2 1.393483e+03
16 UAB003_W7D3 1.065232e+03
17 UAB003_W7D5 7.140310e+02
18 UAB003_W7D6 2.752729e+02
19 UAB003_W7D7 1.561850e+03
20 UAB003_W8D1 5.939187e+02
21 UAB003_W8D2 1.003773e+03
22 UAB005_W4D7 1.816241e+03
23 UAB005_W5D1 1.508920e+03
24 UAB005_W5D2 1.676461e+03
25 UAB005_W5D3 2.577314e+01
26 UAB005_W5D4 1.974515e+03
27 UAB005_W5D5 2.484168e+03
28 UAB006_W3D1 5.126613e+02
29 UAB006_W3D3 1.242925e+03
30 UAB006_W3D4 3.796479e+02
31 UAB006_W3D5 6.575725e+02
32 UAB006_W3D6 9.406057e+02
33 UAB006_W3D7 3.035163e+02
34 UAB006_W4D2 9.091423e+02
35 UAB006_W4D3 1.067289e+03
36 UAB006_W4D4 7.653917e+02
37 UAB006_W4D5 5.108480e+02
38 UAB006_W4D6 1.240235e+00
39 UAB006_W4D7 3.098208e+02
40 UAB006_W5D1 2.580749e+02
41 UAB007_W3D7 1.683195e+03
42 UAB007_W4D1 1.188735e+03
43 UAB007_W4D2 7.899422e+02
44 UAB007_W4D3 3.108406e+02
45 UAB007_W4D4 1.085003e+03
46 UAB008_W10D1 1.179684e+02
47 UAB008_W10D2 2.383264e+02
48 UAB008_W9D5 1.046533e+02
49 UAB008_W9D6 6.643826e+02
50 UAB008_W9D7 9.899674e+02
51 UAB015_W4D3 1.701653e+03
52 UAB015_W4D4 8.764815e+02
53 UAB015_W4D5 2.096003e+02
54 UAB015_W4D7 1.426799e+03
55 UAB015_W5D1 2.021888e+03
56 UAB015_W5D2 4.693895e+02
57 UAB015_W5D3 6.060877e+02
58 UAB015_W5D4 3.483291e+02
59 UAB015_W5D5 1.305746e+02
60 UAB015_W5D7 6.715677e+02
61 UAB015_W6D1 2.496645e+02
62 UAB015_W6D2 4.994072e+02
63 UAB015_W6D3 6.873087e+02

64 UAB015_W6D4 6.197250e+02
65 UAB021_W1D4 3.699508e+02
66 UAB021_W1D5 4.337148e+02
67 UAB021_W1D7 1.449508e+02
68 UAB021_W2D2 1.139957e+02
69 UAB022_W7D5 6.055661e+02
70 UAB022_W7D6 1.071701e+02
71 UAB022_W7D7 1.095501e+03
72 UAB022_W8D1 2.092547e+02
73 UAB022_W8D2 9.373751e+02
74 UAB055_W2D1 7.657517e+02
75 UAB055_W2D2 5.755693e+02
76 UAB055_W2D3 3.015522e+02
77 UAB055_W2D4 1.686811e+02
78 UAB055_W2D5 1.282814e+02
79 UAB055_W2D6 4.453451e+01
80 UAB093_W6D5 1.055765e+03
81 UAB093_W6D6 3.812033e+02
82 UAB093_W6D7 1.968737e+02
83 UAB102_W4D6 8.767737e+01
84 UAB102_W5D1 1.119551e+02
85 UAB102_W5D2 2.917887e+02
86 UAB115_W3D7 4.359042e+02
87 UAB115_W4D3 1.105002e+03
88 UAB115_W4D4 8.179583e+02
89 UAB115_W4D5 1.705125e+03
90 UAB115_W4D6 2.974179e+02
91 UAB115_W5D2 1.155358e+03
92 UAB115_W5D3 7.861445e+02
93 UAB115_W5D5 1.667892e+02
94 UAB115_W5D7 1.417144e+03
95 UAB116_W2D5 1.440792e+02
96 UAB116_W2D6 1.961594e+02
97 UAB116_W3D1 2.916645e+02
98 UAB116_W3D2 3.231621e+02
99 UAB116_W3D3 8.325147e+01
100 UAB117_W7D1 1.419449e+02
101 UAB117_W7D2 3.295152e+02
102 UAB117_W7D3 8.492167e+02
103 UAB117_W7D4 1.340304e+03
104 UAB121_W3D3 5.511017e+02
105 UAB121_W3D4 1.048689e-01
106 UAB121_W3D5 3.092622e+02
107 UAB121_W3D6 1.049983e+03
108 UAB121_W3D7 2.265420e+02
109 UAB121_W4D2 1.407170e-01
110 UAB121_W4D3 8.314880e+02
111 UAB121_W4D4 8.507223e-02
112 UAB121_W4D5 2.409049e+03
113 UAB121_W4D6 9.106602e+02

```
hist(log(colSums(SRL_counts_table)/number_mirs_genome,10),main="Log10 Total
miRNA Counts/# miRNAs")
abline(v = log(100,10))
```



```
.1*number_mirs_genome

## Features
## 186.9

## Samples with "low coverage", i.e., less than 1E4 reads/#miRs
low_coverage<-(colSums(SRL_counts_table))<=(125)*number_mirs_genome ##125
sum(low_coverage)

## [1] 13

125 * 1869

## [1] 233625

## An idea of the samples with "low coverage":
print(head(SRL_counts_table[,low_coverage]))

##          UAB005_W5D3 UAB006_W4D6 UAB008_W10D1 UAB008_W9D5 UAB021_W2D2
## hsa-let-7a-2      602         36         2241         2463         2788
## hsa-let-7a-3      618         37         2288         2529         2905
```

```

## hsa-let-7b          1868          229          7186          11804          17191
## hsa-let-7c           657           61          2766           5179           4186
## hsa-let-7d           98            1           271            262            338
## hsa-let-7e           10            2            88             81             36
##                   UAB022_W7D6 UAB055_W2D6 UAB102_W4D6 UAB102_W5D1 UAB116_W3D3
## hsa-let-7a-2        3438          1652          2340          2646          1296
## hsa-let-7a-3        3531          1621          2487          2713          1302
## hsa-let-7b         20351          6993          19034          27855          11430
## hsa-let-7c         11309          3736          6052          12083          5938
## hsa-let-7d           792           206           305            413            180
## hsa-let-7e           161            48            54             64             33
##                   UAB121_W3D4 UAB121_W4D2 UAB121_W4D4
## hsa-let-7a-2         10            9             10
## hsa-let-7a-3         9             15            10
## hsa-let-7b           25            23            10
## hsa-let-7c           5             9             10
## hsa-let-7d           0             1             0
## hsa-let-7e           0             0             0

```

```

(remove_samples<-remove_poorQC_samples(remove_samples =
remove_samples,sample_list
=unique(colnames(SRL_counts_table[,low_coverage])),reason = "Low_Coverage"))

```

```

##       Pre_QC_ID QC_removal_stage
## 1  UAB005_W5D3   Low_Coverage
## 2  UAB006_W4D6   Low_Coverage
## 3  UAB008_W10D1  Low_Coverage
## 4  UAB008_W9D5   Low_Coverage
## 5  UAB021_W2D2   Low_Coverage
## 6  UAB022_W7D6   Low_Coverage
## 7  UAB055_W2D6   Low_Coverage
## 8  UAB102_W4D6   Low_Coverage
## 9  UAB102_W5D1   Low_Coverage
## 10 UAB116_W3D3   Low_Coverage
## 11 UAB121_W3D4   Low_Coverage
## 12 UAB121_W4D2   Low_Coverage
## 13 UAB121_W4D4   Low_Coverage

```

```

## Remove those samples with low coverage
counts_meta.qc<-subset_ExpressionSet(expSet = SRL_counts_meta,filterOut =
c(as.character(remove_samples$Pre_QC_ID)))

```

```
nrow(pData(counts_meta.qc))
```

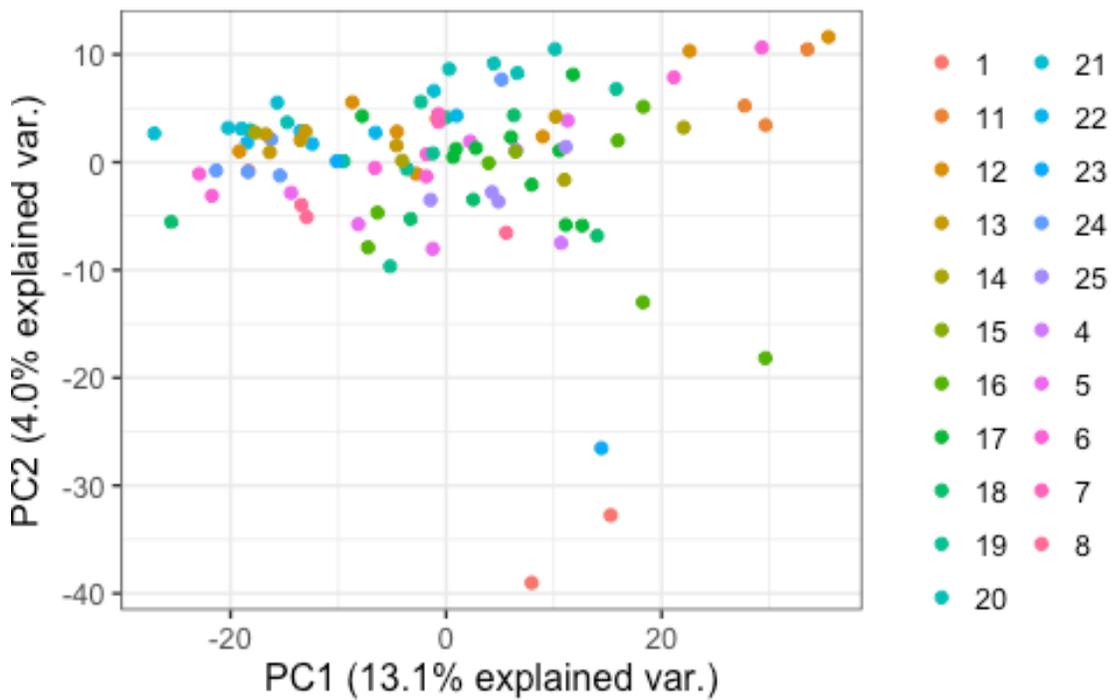
```
## [1] 100
```

```
ncol(exprs(counts_meta.qc))
```

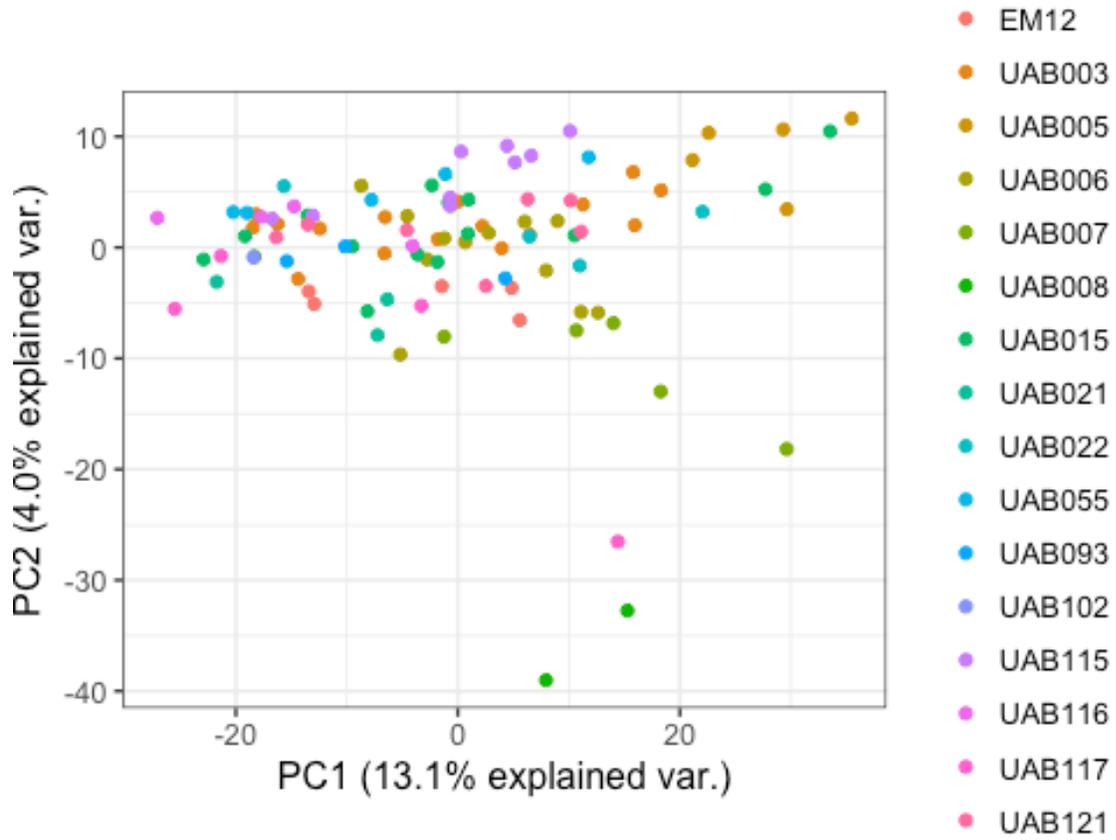
```
## [1] 100
```

```
ncol(pData(counts_meta.qc))
```

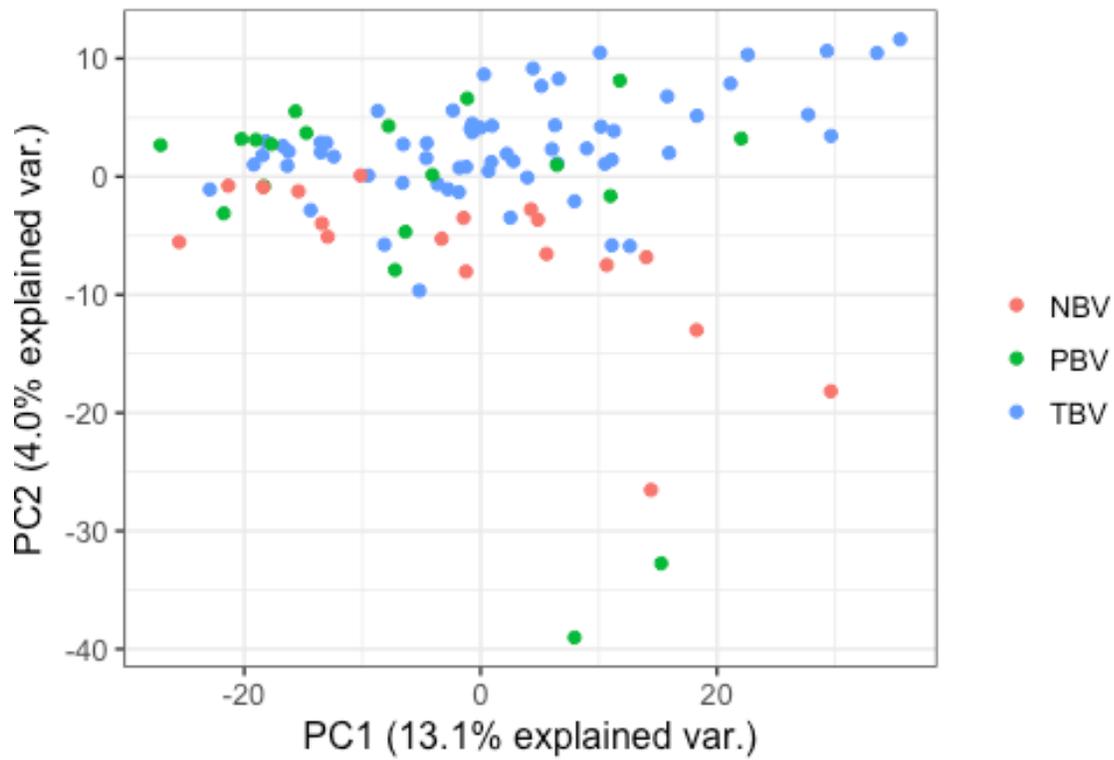
```
## [1] 50
nrow(exprs(counts_meta.qc))
## [1] 1546
## Re-plot after QC
plot_QC.postQC.batch<-plot_pca(counts_meta.qc,"Log miRNA Raw Count Distance,
POST QC",color_by = "Batch",ploly = F,seed_val = seed_val)##FIGURE
plot_QC.postQC.batch$pca.p
```



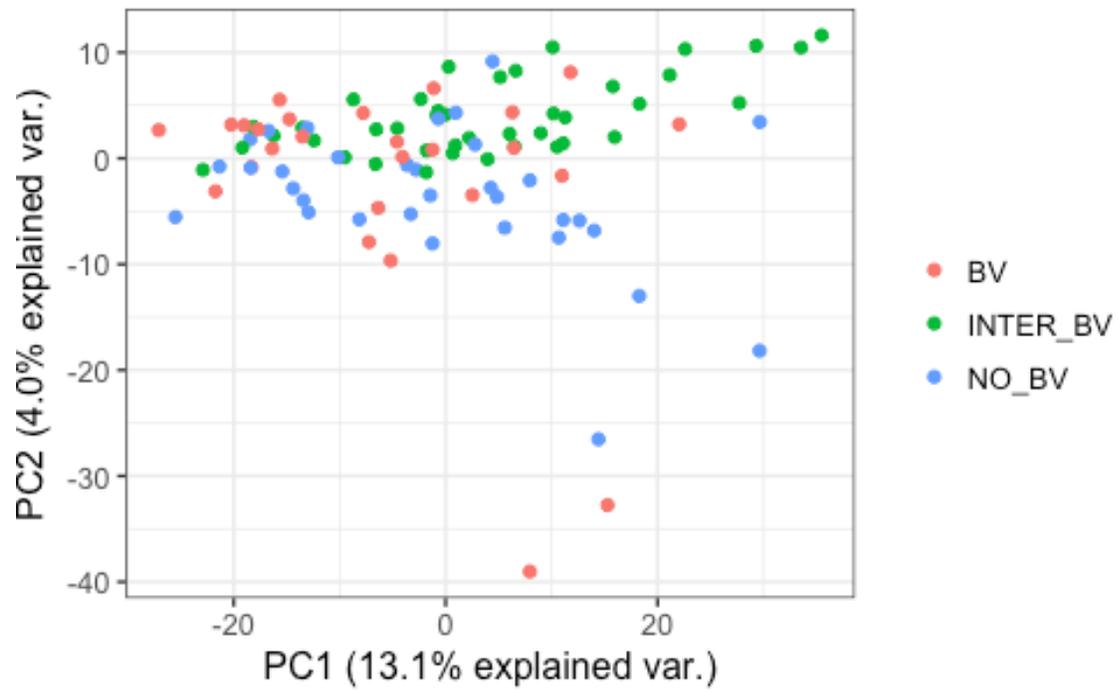
```
plot_QC.postQC.sid<-plot_pca(counts_meta.qc,"Log miRNA Raw Count Distance,
POST QC",color_by = "SID",ploly = F,seed_val = seed_val)##FIGURE
plot_QC.postQC.sid$pca.p
```



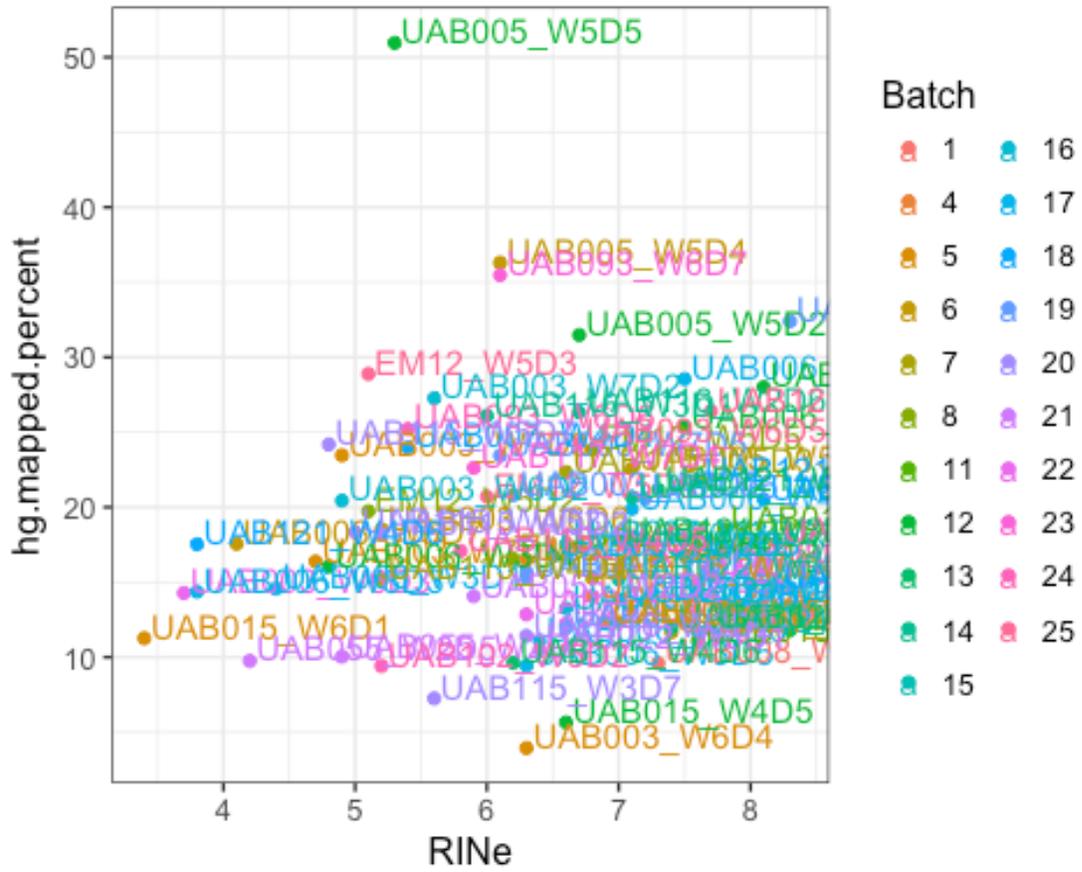
```
plot_pca(counts_meta.qc, "Log miRNA Raw Count Distance, POST QC", color_by =
"BVGroup", poly = F, seed_val = seed_val)$pca.p##FIGURE
```



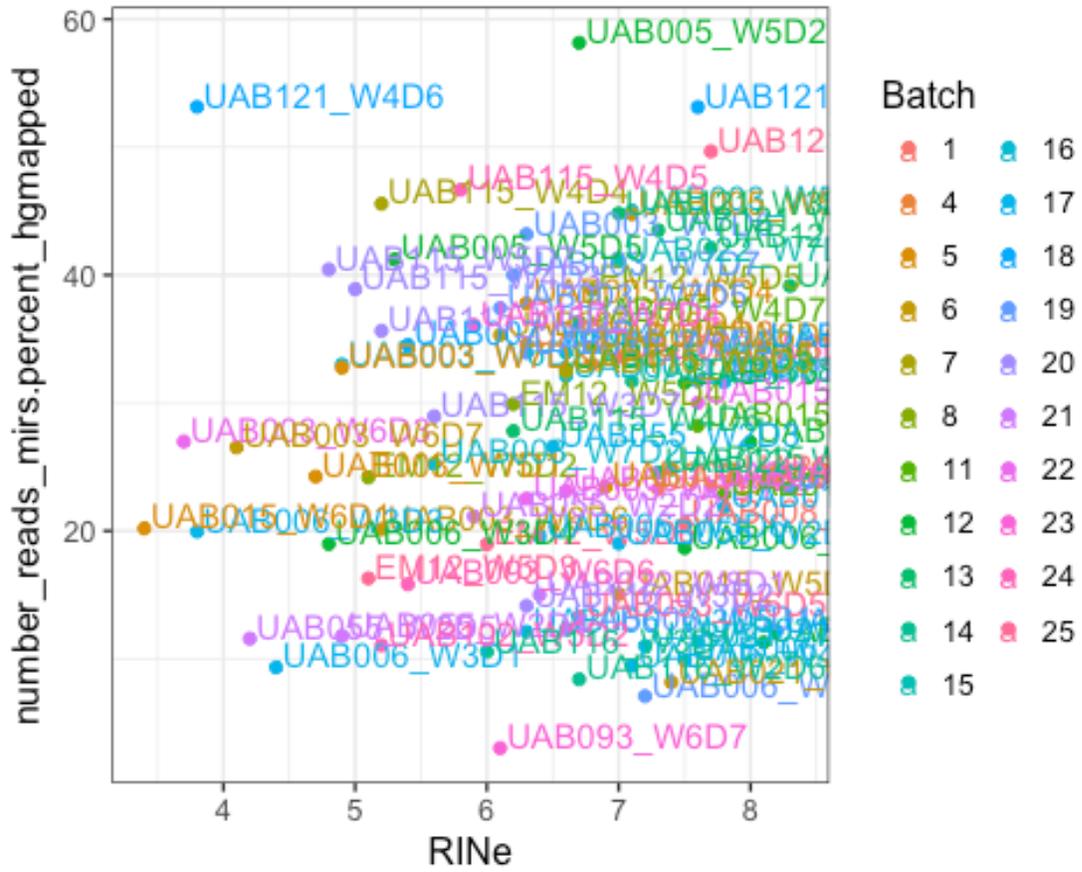
```
plot_pca(counts_meta.qc, "Log miRNA Raw Count Distance, POST QC", color_by =  
"NugentC", polyl = F, seed_val = seed_val)$pca.p##FIGURE
```



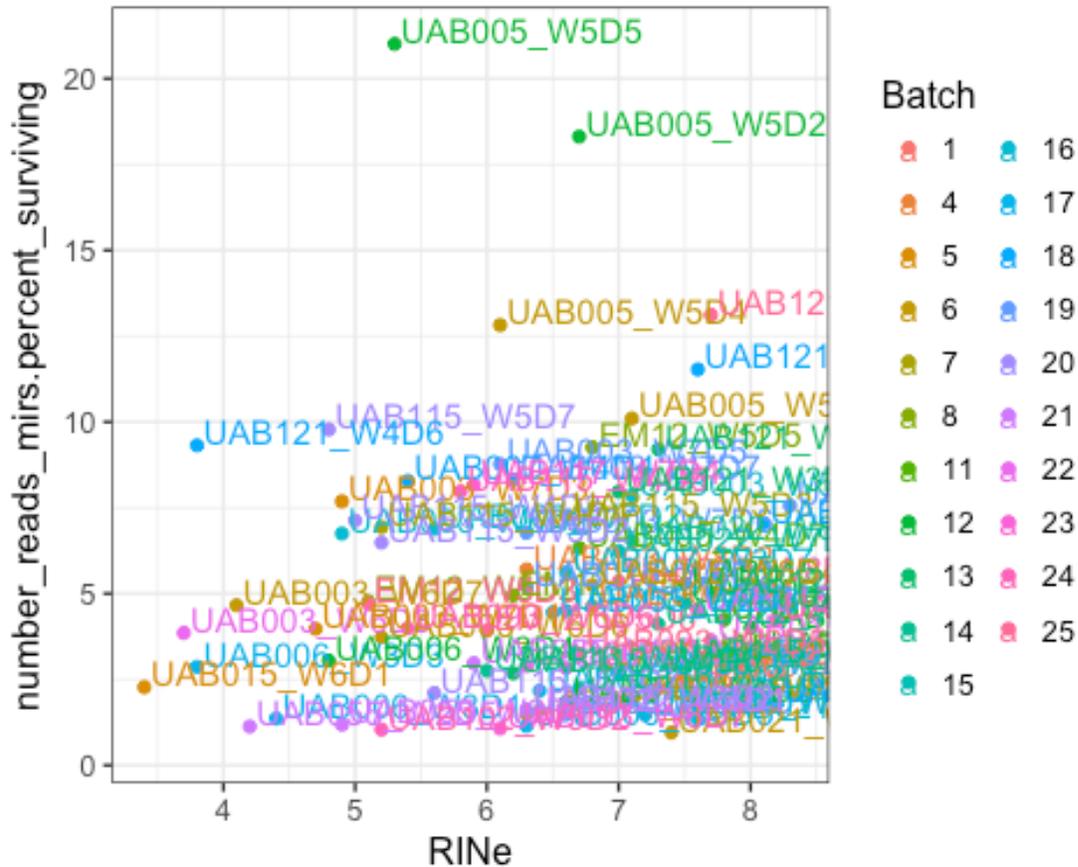
```
plot_RIN_meta(pData(counts_meta.qc),"hg.mapped.percent",col_by =  
"Batch")##FIGURE
```



```
plot_RIN_meta(pData(counts_meta.qc), "number_reads_mirs.percent_hgmapped", col_
by = "Batch")##FIGURE
```



```
plot_RIN_meta(pData(counts_meta.qc),"number_reads_mirs.percent_surviving",col
_by = "Batch")##FIGURE
```



```
dev.off()

## null device
##          1

median_RIN<-median(pData(counts_meta.qc)$RIN,na.rm = T)

print(paste0("The median RINe score was ",median_RIN," with
",round(100*sum(pData((counts_meta.qc))$RIN<7,na.rm =
T)/length(pData((counts_meta.qc))$RIN),digits = 1),"% of the samples having a
RINe greater than 7."))

## [1] "The median RINe score was 6.6, with 60% of the samples having a RINe
greater than 7."

number_miRNAs<-mapping_stats(pData(counts_meta.qc)$number_reads_mirs)
percent_miRNAs.hg<-
round(mapping_stats(pData(counts_meta.qc)$number_reads_mirs.percent_hgmapped)
,1)
percent_miRNAs.oftrimmed<-
round(mapping_stats(pData(counts_meta.qc)$number_reads_mirs.percent_surviving
),1)

print(paste0("The median (minimum/maximum) percentage of post-QC miRNA reads
```

```

relative to all hg19 mapped and total reads was ",percent_miRNAs.hg$median,"%
(",percent_miRNAs.hg$min,"%/",percent_miRNAs.hg$max,"%") and
",percent_miRNAs.oftrimmed$median,"%
(",percent_miRNAs.oftrimmed$min,"%/",percent_miRNAs.oftrimmed$max,"%"),
respectively. However, the median number of post-QC hg19 mapped miRNA reads
was ",number_miRNAs$median,", with a minimum ",number_miRNAs$min," and
maximum ",number_miRNAs$max,". Thus, despite low relative miRNA read counts,
the estimated coverage ranged from
",round(number_miRNAs$min/number_mirs_genome,digits = 0),"X-
",round(number_miRNAs$max/number_mirs_genome,digits = 0),"X across the entire
miRnome ("",number_mirs_genome," annotated miRNAs).")

```

```

## [1] "The median (minimum/maximum) percentage of post-QC miRNA reads
relative to all hg19 mapped and total reads was 26.6% (3%/58.2%) and 4.2%
(0.5%/21%), respectively. However, the median number of post-QC hg19 mapped
miRNA reads was 1204913, with a minimum 239758 and maximum 4642910. Thus,
despite low relative miRNA read counts, the estimated coverage ranged from
128X-2484X across the entire miRnome (1869 annotated miRNAs)."

```

```

sum(pData((counts_meta.qc))$number_reads_mirs<10^6,na.rm =
T)/length(pData((counts_meta.qc))$number_reads_mirs)

```

```
## [1] 0.39
```

```

cairo_ps(file =
paste0(thesis_figures_directory,FIGURE_QC_PCA.PREQC.BYBATCH),width = 8,height
= 5.5)
plot(plot_QC.preQC.batch$pca.p)
dev.off()

```

```
## null device
##          1
```

```

cairo_ps(file =
paste0(thesis_figures_directory,FIGURE_QC_PCA.PREQC.BYSUBJ),width = 8,height
= 5.5)
plot(plot_QC.preQC.sid$pca.p)
dev.off()

```

```
## null device
##          1
```

```

cairo_ps(file =
paste0(thesis_figures_directory,FIGURE_QC_PCA.RMLOW.BYBATCH),width = 8,height
= 5.5)
plot(plot_QC.postQC.batch$pca.p)
dev.off()

```

```
## null device
##          1
```

```

cairo_ps(file =
paste0(thesis_figures_directory,FIGURE_QC_PCA.RMLOW.BYSUBJ),width = 8,height
= 5.5)
plot(plot_QC.postQC.sid$pca.p)
dev.off()

## null device
##          1

cairo_ps(file =
paste0(thesis_figures_directory,FIGURE_QC_RIN_v_READS.PROP),width = 8,height
= 5.5)
plot(plot_QC.RIN.percent.batch)
dev.off()

## null device
##          1

cairo_ps(file =
paste0(thesis_figures_directory,FIGURE_QC_RIN_v_READS.ABS),width = 8,height =
5.5)
plot(plot_QC.RIN.number.batch)
dev.off()

## null device
##          1

```

Discovery of miRNAs Associated w/ Lactobacillus spp. vs CST-IV/BV

Build a proxy-Amsel model from Clinical Visits

Not all samples in the miRNA cohort have Amsel scores, but they do have 16S rRNA sequencing taxa assignments. Therefore:

1. Build a RF model that uses a separate cohort to predict Amsel-16S rRNA associations.
2. Assess accuracy, and tweak parameters or predictors.
3. Using this model, assign proxy-Amsel scores to the miRNA cohort from the available 16S rRNA data.

```

## Generate a model using clinical visit data. The model uses 16S relative
abundances to predict Amsel diagnosis.
load(file=paste0(R_script_input_directory,"CV_16S_AMSEL.Rdata"))
load(file=paste0(R_script_input_directory,"SRL_16S.Rdata"))
SRL_16S<-SRL_16S[!row.names(SRL_16S) %in%
as.character(removed_samples$Pre_QC_ID),]

ncol(CV_16S_AMSEL)

## [1] 204

setdiff(names(CV_16S_AMSEL),names(SRL_16S))

```

```

## [1] "AMSEL"      "Amsel_BV"

setdiff(names(SRL_16S),names(CV_16S_AMSEL))

## character(0)

table(CV_16S_AMSEL$Amsel_BV)

##
## NBV PBV
## 210  71

nuniq_subjs_amsel<-length(unique(gsub(row.names(CV_16S_AMSEL),pattern =
"_.*",replacement = "")))

print(paste0('Therefore, a proxy-Amsel diagnosis was developed by applying a
Random Forest model trained with metataxonomic data and metadata from
',nuniq_subjs_amsel,' subjects of the parent cohort that included
',nrow(CV_16S_AMSEL),' samples for which both metataxonomic data and Amsel
diagnosis was available (Amsel subset)(',TABLE_PROXY_AMSEL_INPUT,').'))

## [1] "Therefore, a proxy-Amsel diagnosis was developed by applying a Random
Forest model trained with metataxonomic data and metadata from 117 subjects
of the parent cohort that included 281 samples for which both metataxonomic
data and Amsel diagnosis was available (Amsel subset)(TABLE_A7.csv)."
```

```

NBV.amsel<-table(CV_16S_AMSEL$Amsel_BV)["NBV"]
PBV.amsel<-table(CV_16S_AMSEL$Amsel_BV)["PBV"]

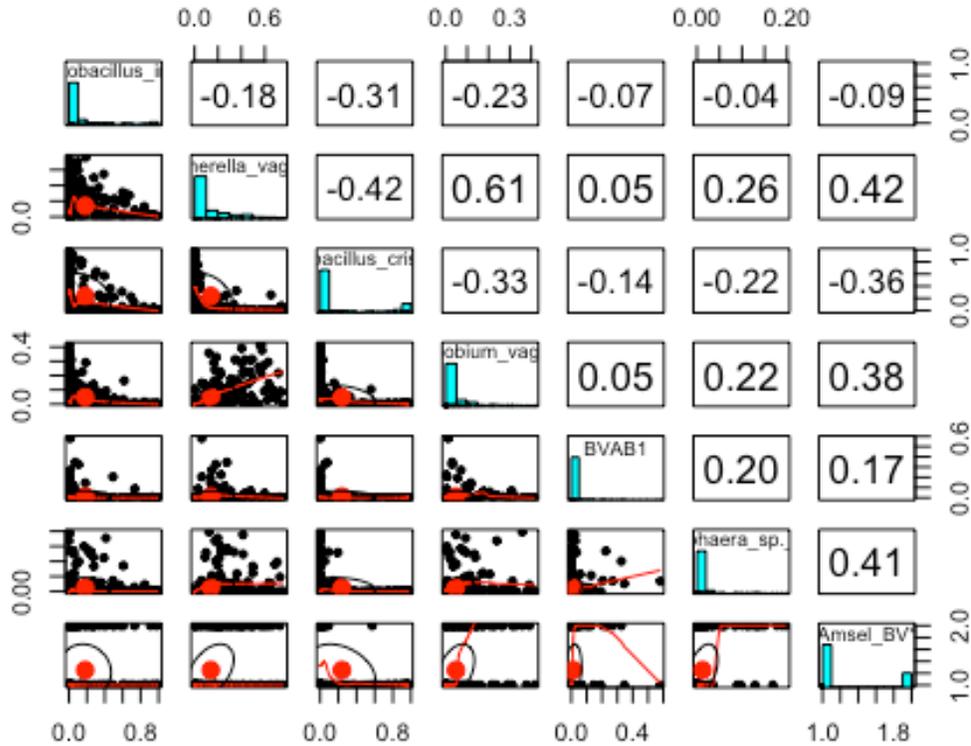
print(paste0("In the Amsel subset, asymptomatic or symptomatic BV diagnoses
represented ",round(100*PBV.amsel/(NBV.amsel+PBV.amsel),digits = 1),"%
(",PBV.amsel,"/",NBV.amsel+PBV.amsel,") of the Amsel diagnoses
(",TABLE_PROXY_AMSEL_INPUT,") a figure closely matching the reported
prevalence of 29.2% for BV in similar populations [REF]."))

## [1] "In the Amsel subset, asymptomatic or symptomatic BV diagnoses
represented 25.3% (71/281) of the Amsel diagnoses (TABLE_A7.csv), a figure
closely matching the reported prevalence of 29.2% for BV in similar
populations [REF]."
```

```

##Sanity check for Amsel
pairs.panels(dplyr::select(CV_16S_AMSEL,Lactobacillus_iners,Gardnerella_vagin
alis,Lactobacillus_crispatus,Atopobium_vaginae,BVAB1,Megasphaera_sp._type_1,A
msel_BV),scale = F,density = F)

```



```
CV_16S_AMSEL <- CV_16S_AMSEL %>% dplyr::select(-c(AMSEL))
min_CV_16S_AMSEL<-min(dplyr::select(CV_16S_AMSEL,-
Amsel_BV)[dplyr::select(CV_16S_AMSEL,-Amsel_BV)>0],na.rm = T)
CV_16S_AMSEL[is.na(CV_16S_AMSEL)]<-0
CV_16S_AMSEL_IDs=row.names(CV_16S_AMSEL)
```

See run_randomForest function above for paramaters. Note that you can load a previously-built model to save time.

```
Clinical_Visit_RF<-run_randomForest(predictors_response_table = CV_16S_AMSEL,
response_variable_name = "Amsel_BV",
subj_spec = FALSE,
nfold = nfolds,
nreps = npermutes,
save_model = F,
file_n = "Clinical_Visit_RF",
verbose = F,
permute = T,
load_prev_model = T,
```

```
R_script_output_directory=R_script_output_directory)
```

```

Clinical_Visit_RF$accuracy_table

##      Var2
## Var1 NBV PBV
## NBV  55  5
## PBV   8 16

(Clinical_Visit_RF.accuracy<-
c(diag(Clinical_Visit_RF$accuracy_table)/rowSums(Clinical_Visit_RF$accuracy_t
able)))

##      NBV      PBV
## 0.9166667 0.6666667

(Clinical_Visit_RF$top_features$top_features.all<-
row.names(Clinical_Visit_RF$importance[Clinical_Visit_RF$importance$MeanDecre
aseAccuracy.pval<=pval_threshold |
Clinical_Visit_RF$importance$MeanDecreaseGini.pval<=pval_threshold,]))

## [1] "Parvimonas_micra"          "Megasphaera_sp._type_1"
## [3] "BVAB2"                        "Eggerthella"
## [5] "Dialister_sp._type_2"         "Leptotrichia_amnionii"
## [7] "BVAB3"                        "BVAB1"
## [9] "PH"                            "Prevotella_genogroup_3"
## [11] "Prevotella_genogroup_1"       "Mobiluncus_mulieris"
## [13] "Porphyromonas_sp._type_1"     "Gardnerella_vaginalis"
## [15] "Prevotella_genogroup_2"       "Peptoniphilus_lacrimalis"
## [17] "Atopobium_vaginae"           "MENSTRUATION_NORMALIZED_PHASED"
## [19] "Prevotella_genogroup_4"       "Candidate_Division_TM7_vaginal"
## [21] "Lactobacillus_crispatus"      "Prevotella_genogroup_5"
## [23] "Porphyromonas_uenonis"        "Gemella"
## [25] "Anaerococcus_vaginalis"       "Lactobacillus_helveticus"
## [27] "Prevotella_melaninogenica"    "Lactobacillus_vaginalis"
## [29] "Porphyromonas_endodontalis"   "Sutterella_stercoricanis"
## [31] "Firmicutes"                   "Anaerococcus"

print(paste0("Multiple
(",length(Clinical_Visit_RF$top_features$top_features.all),") important
microbial (taxa and their relative abundance) or metadata features were
predictive of the Amsel diagnosis:
",str_c(sort(Clinical_Visit_RF$top_features$top_features.all),collapse = ",
")," (" ,TABLE_RF_SUMMARY.CV,"))

## [1] "Multiple (32) important microbial (taxa and their relative abundance)
or metadata features were predictive of the Amsel diagnosis: Anaerococcus,
Anaerococcus_vaginalis, Atopobium_vaginae, BVAB1, BVAB2, BVAB3,
Candidate_Division_TM7_vaginal, Dialister_sp._type_2, Eggerthella,
Firmicutes, Gardnerella_vaginalis, Gemella, Lactobacillus_crispatus,
Lactobacillus_helveticus, Lactobacillus_vaginalis, Leptotrichia_amnionii,
Megasphaera_sp._type_1, MENSTRUATION_NORMALIZED_PHASED, Mobiluncus_mulieris,

```

```
Parvimonas_micra, Peptoniphilus_lacrimalis, PH, Porphyromonas_endodontalis,
Porphyromonas_sp._type_1, Porphyromonas_uenonis, Prevotella_genogroup_1,
Prevotella_genogroup_2, Prevotella_genogroup_3, Prevotella_genogroup_4,
Prevotella_genogroup_5, Prevotella_melaninogenica, Sutterella_stercoricanis
(TABLE_A8.csv)"
```

```
print(paste0("The Amsel Random Forest model accuracy was tested using a hold-
out set and found to be ",round(100*Clinical_Visit_RF.accuracy["NBV"],1),"%
accurate in correctly assigning NBV diagnosis and
",round(100*Clinical_Visit_RF.accuracy["PBV"],1),"% accurate in correctly
assigning PBV diagnosis."))
```

```
## [1] "The Amsel Random Forest model accuracy was tested using a hold-out
set and found to be 91.7% accurate in correctly assigning NBV diagnosis and
66.7% accurate in correctly assigning PBV diagnosis."
```

```
proxy_amsel_predictions.classprobs<-predict(Clinical_Visit_RF$rfp, SRL_16S,
type = "prob")
(proxy_amsel_predictions<-predict(Clinical_Visit_RF$rfp, SRL_16S, type =
"response"))
```

```
## UAB003_W5D6 UAB003_W6D1 UAB003_W6D2 UAB003_W6D3 UAB003_W6D4 UAB003_W6D5
## NBV NBV NBV NBV NBV NBV
## UAB003_W6D6 UAB003_W6D7 UAB003_W7D1 UAB003_W7D2 UAB003_W7D3 UAB003_W7D5
## NBV NBV NBV NBV NBV NBV
## UAB003_W7D6 UAB003_W7D7 UAB003_W8D1 UAB003_W8D2 UAB005_W4D7 UAB005_W5D1
## NBV NBV NBV NBV NBV NBV
## UAB005_W5D2 UAB005_W5D4 UAB005_W5D5 UAB006_W3D3 UAB006_W3D4 UAB006_W3D5
## NBV NBV NBV NBV NBV NBV
## UAB006_W3D6 UAB006_W3D7 UAB006_W4D2 UAB006_W4D3 UAB006_W4D4 UAB006_W4D5
## NBV NBV PBV PBV PBV PBV
## UAB006_W4D7 UAB006_W5D1 UAB007_W3D7 UAB007_W4D1 UAB007_W4D2 UAB007_W4D3
## PBV PBV NBV NBV NBV NBV
## UAB007_W4D4 UAB008_W9D6 UAB015_W4D3 UAB015_W4D4 UAB015_W4D5 UAB015_W4D7
## NBV PBV PBV PBV PBV PBV
## UAB015_W5D1 UAB015_W5D2 UAB015_W5D3 UAB015_W5D4 UAB015_W5D5 UAB015_W5D7
## PBV PBV PBV NBV PBV PBV
## UAB015_W6D1 UAB015_W6D2 UAB015_W6D3 UAB015_W6D4 UAB021_W1D5 UAB021_W1D7
## NBV NBV PBV NBV PBV PBV
## UAB022_W8D1 UAB055_W2D1 UAB055_W2D2 UAB055_W2D3 UAB055_W2D4 UAB055_W2D5
## PBV PBV PBV PBV PBV PBV
## UAB093_W6D6 UAB115_W3D7 UAB115_W4D3 UAB115_W4D4 UAB115_W4D6 UAB115_W5D2
## NBV NBV NBV NBV NBV NBV
## UAB115_W5D3 UAB115_W5D5 UAB115_W5D7 UAB116_W2D5 UAB116_W2D6 UAB116_W3D1
## NBV NBV NBV PBV PBV PBV
## UAB116_W3D2 UAB117_W7D2 UAB117_W7D4 UAB121_W3D6 UAB121_W4D3 UAB121_W4D5
## PBV NBV NBV NBV NBV NBV
## EM12_W5D2 EM12_W5D3 EM12_W5D4 EM12_W5D5 EM12_W5D6
## NBV NBV NBV NBV NBV
## Levels: NBV PBV
```

```

CV_16S_AMSEL.out<-CV_16S_AMSEL
CV_16S_AMSEL.out$Amsel_BV_trainingTestingSet<-""
CV_16S_AMSEL.out[row.names(CV_16S_AMSEL.out) %in%
Clinical_Visit_RF$testing_ids,"Amsel_BV_trainingTestingSet"]<-"testing"
CV_16S_AMSEL.out[row.names(CV_16S_AMSEL.out) %in%
Clinical_Visit_RF$training_ids,"Amsel_BV_trainingTestingSet"]<-"training"

write.csv(CV_16S_AMSEL.out,file=paste0(thesis_tables_directory,TABLE_PROXY_AMSEL_INPUT),row.names=T,quote=F)

SRL_16S_AMSEL<-cbind(SRL_16S,AMSEL_PREDICTION=proxy_amsel_predictions)
table(SRL_16S_AMSEL$AMSEL_PREDICTION)

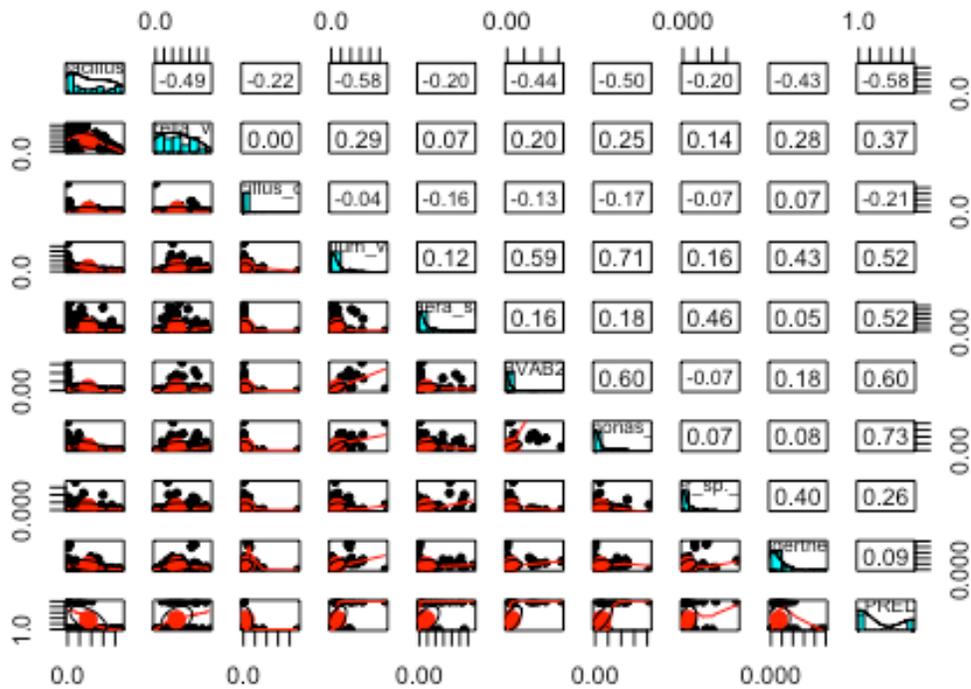
##
## NBV PBV
## 54 29

write.csv(SRL_16S_AMSEL,file=paste0(thesis_tables_directory,TABLE_PROXY_AMSEL_SRL),row.names=T,quote=F)

##Sanity Check
nfeats_plot<-9
pairs_panel_data<-SRL_16S_AMSEL[,names(SRL_16S_AMSEL) %in%
c(unique(c("Lactobacillus_iners","Gardnerella_vaginalis","Lactobacillus_crispatus","Atopobium_vaginae",Clinical_Visit_RF$top_features$top_features.all)[1:nfeats_plot]),"AMSEL_PREDICTION")]
pairs.panels(pairs_panel_data,main="proxy-Amsel Prediction Top Features with Gvag, Lcrisp, Avag, Liners")

```

msel Prediction Top Features with Gvag, Lcrisp, Avaq



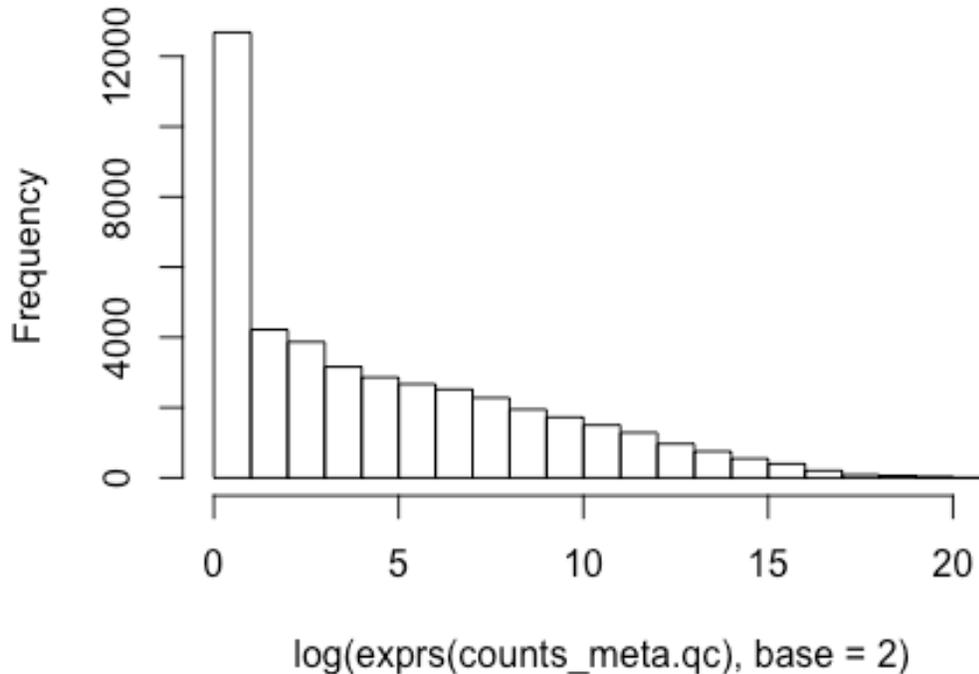
```
dev.off()
```

```
## null device
##          1
```

Normalize & log2 transform small RNA-Seq counts for use in models

```
## Calculte size factors for normalization using calcNormFactors function
hist(log(exprs(counts_meta.qc),base = 2))
```

Histogram of $\log(\text{exprs}(\text{counts_meta.qc}), \text{base} = 2)$



```
expSet.sizeDisp<-calcNormFactors(method = "TMM",DGEList(counts =
exprs(counts_meta.qc)),logratioTrim = 0.10,sumTrim = 0.05)

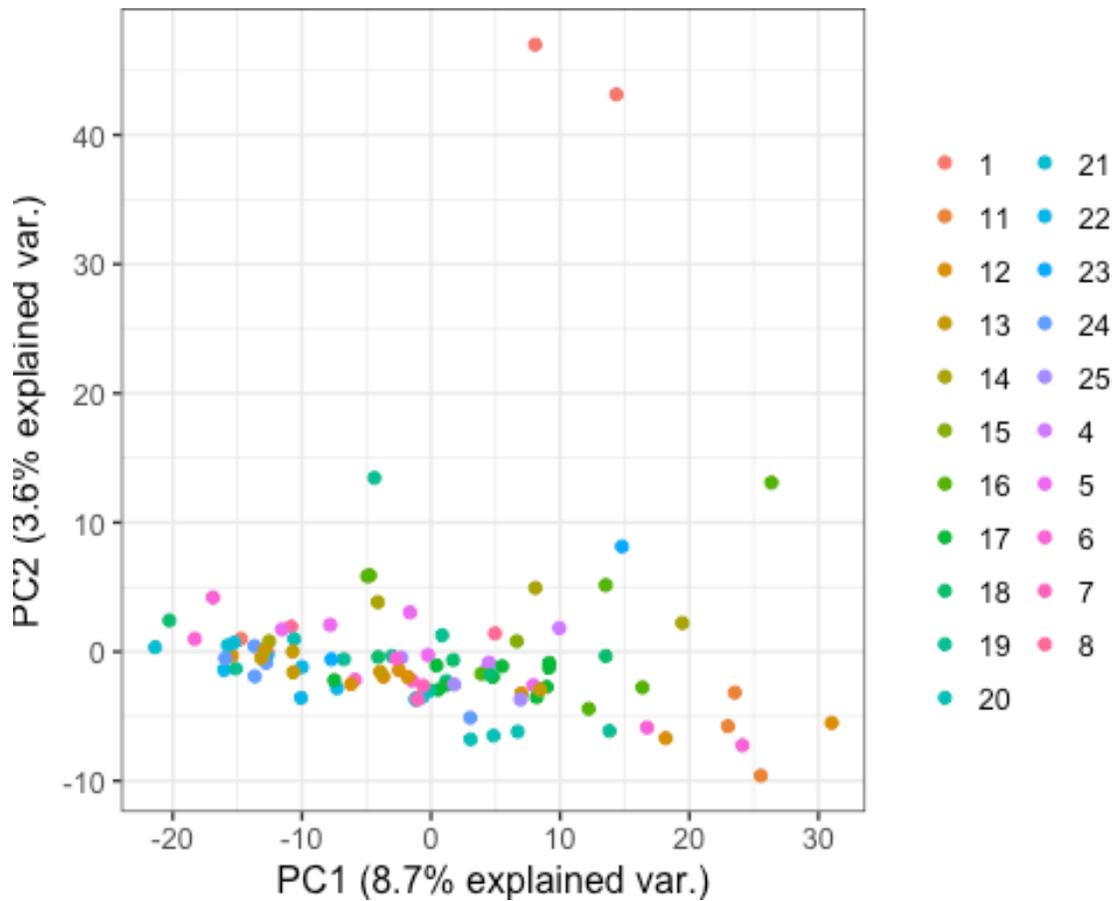
## Normalize counts using size dispersions and norm factors, store as
Expression Set
expSet.normalized<-ExpressionSet(assayData =
as.matrix(1E6*expSet.sizeDisp$counts*expSet.sizeDisp$samples$norm.factors/exp
Set.sizeDisp$samples$lib.size),phenoData =
AnnotatedDataFrame(pData(counts_meta.qc)))

## Store log-transformed QC *NON NORMALIZED* count table for PCA. This is not
used for anything but PCA.
expSet_log<-ExpressionSet(assayData =
log(exprs(counts_meta.qc),base=2),phenoData =
AnnotatedDataFrame(pData(counts_meta.qc)))

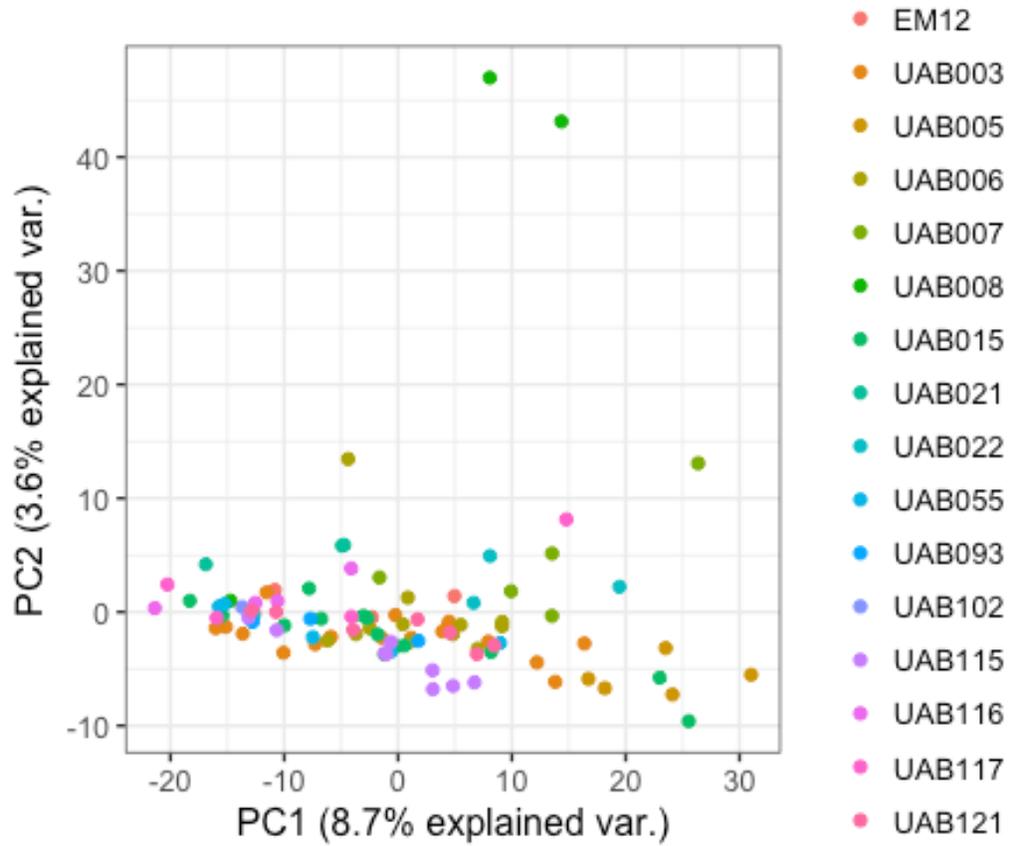
## Store log-transformed QC normalized count table. This is the small RNAseq
count table used in all downstream analysis
expSet_log.normalized<-ExpressionSet(assayData =
log(exprs(expSet.normalized)+1,base=2),
                                phenoData =
AnnotatedDataFrame(pData(expSet.normalized))) ## add pseudo counts... so that
when take log(), all 0 reads become 0 normalized reads log(raw + 1)=log(1)=0.
```

Little effect on other counts. This makes it easier to drop raw 0 counts later. Also store count table annotation from original Expression Set

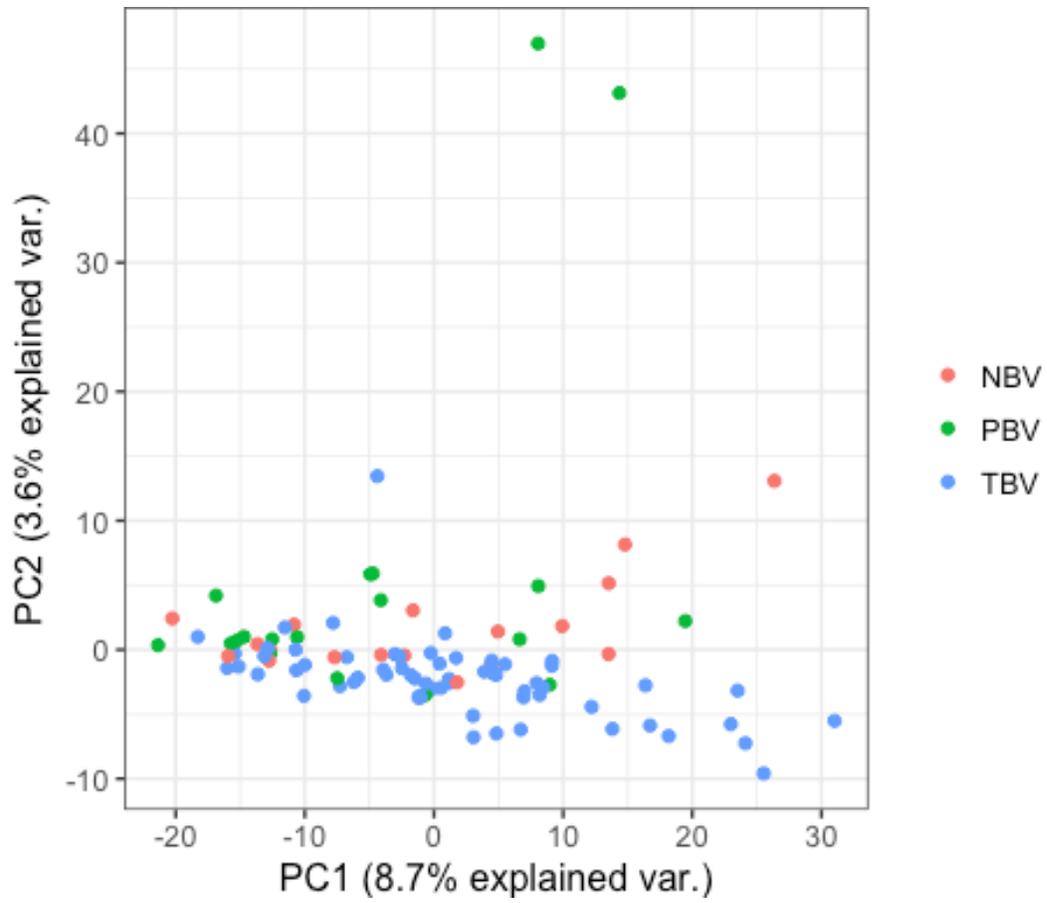
```
## PCA plots after normalization, colored by batch, SID and "BV group"  
plot_QC.normalized.batch<-plot_pca(expSet.normalized,color_by =  
"Batch",margins = unit(c(0,-200,0,-200),units = "points"),seed_val =  
seed_val)  
plot_QC.normalized.batch$pca.p
```



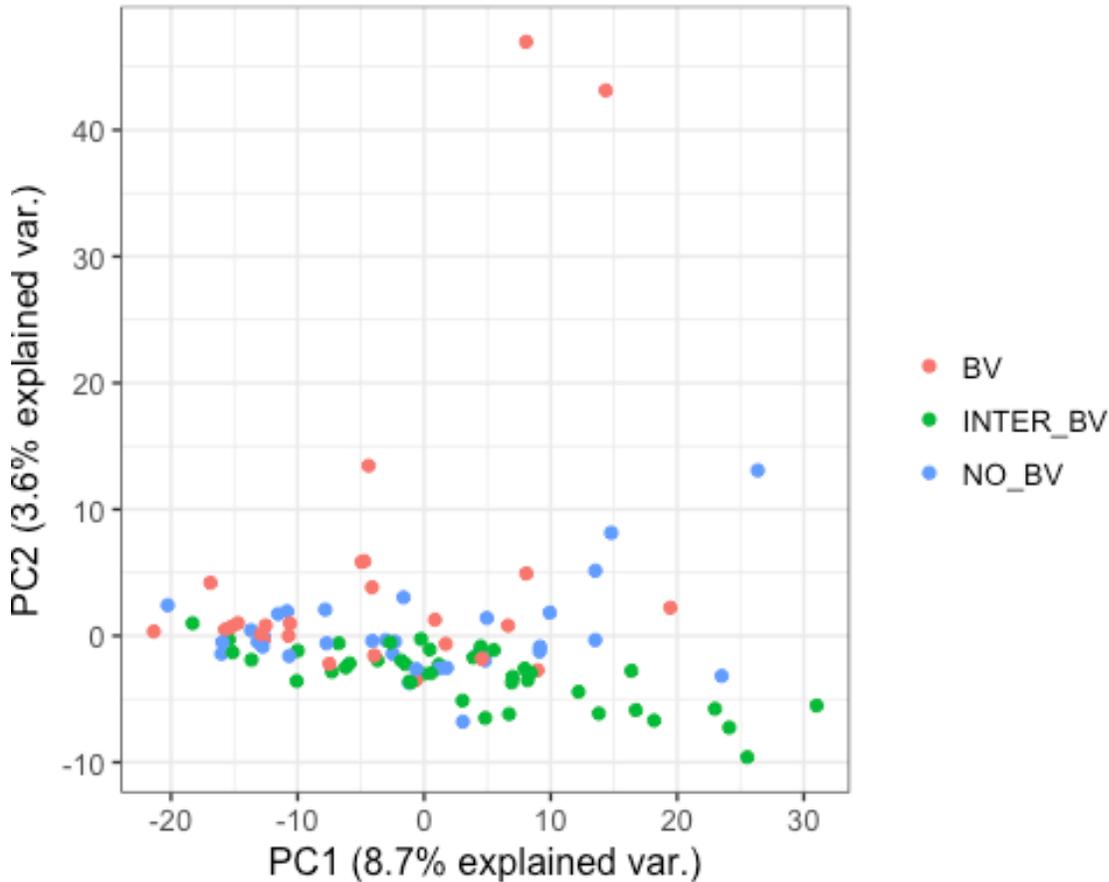
```
plot_QC.normalized.sid<-plot_pca(expSet_log.normalized,color_by = "SID",polly  
= F,logt = F,margins = unit(c(0,30,0,5),units = "pt"),seed_val = seed_val)  
plot_QC.normalized.sid$pca.p
```



```
plot_pca(expSet_log.normalized,color_by = "BVGroup",logt = F)$pca.p
```



```
plot_pca(expSet_log.normalized,color_by = "NugentC",logt = F)$pca.p
```



```
## Print for text
print(paste0("PCA plots before and after sample removal and after
normalization do not support batch effects (guided PCA p-values
",plot_QC.preQC.batch$gPCA.result$p.val,"
",plot_QC.postQC.batch$gPCA.result$p.val," and
",plot_QC.normalized.batch$gPCA.result$p.val," respectively), or subject-
specific effects (guided PCA p-values
",plot_QC.preQC.sid$gPCA.result$p.val,"
",plot_QC.postQC.sid$gPCA.result$p.val," and
",plot_QC.normalized.sid$gPCA.result$p.val," respectively, Figure S1)."))

## [1] "PCA plots before and after sample removal and after normalization do
not support batch effects (guided PCA p-values 0.994, 0.267, and 0.124,
respectively), or subject-specific effects (guided PCA p-values 0.288, 0.478,
and 0.294, respectively, Figure S1)."
```

```
## Store counts data as "features"
features<-exprs(expSet_log.normalized)
head(features)[,1:3]

##           EM12_W5D2 EM12_W5D3 EM12_W5D4
## hsa-let-7a-2 14.554033 13.956142 13.936057
## hsa-let-7a-3 14.750121 14.559435 15.451481
## hsa-let-7b   16.993347 19.089923 18.246961
```

```

## hsa-let-7c    15.085236 14.779020 15.343028
## hsa-let-7d    10.933057 10.523169 12.401490
## hsa-let-7e     8.730594  9.655014  9.324838

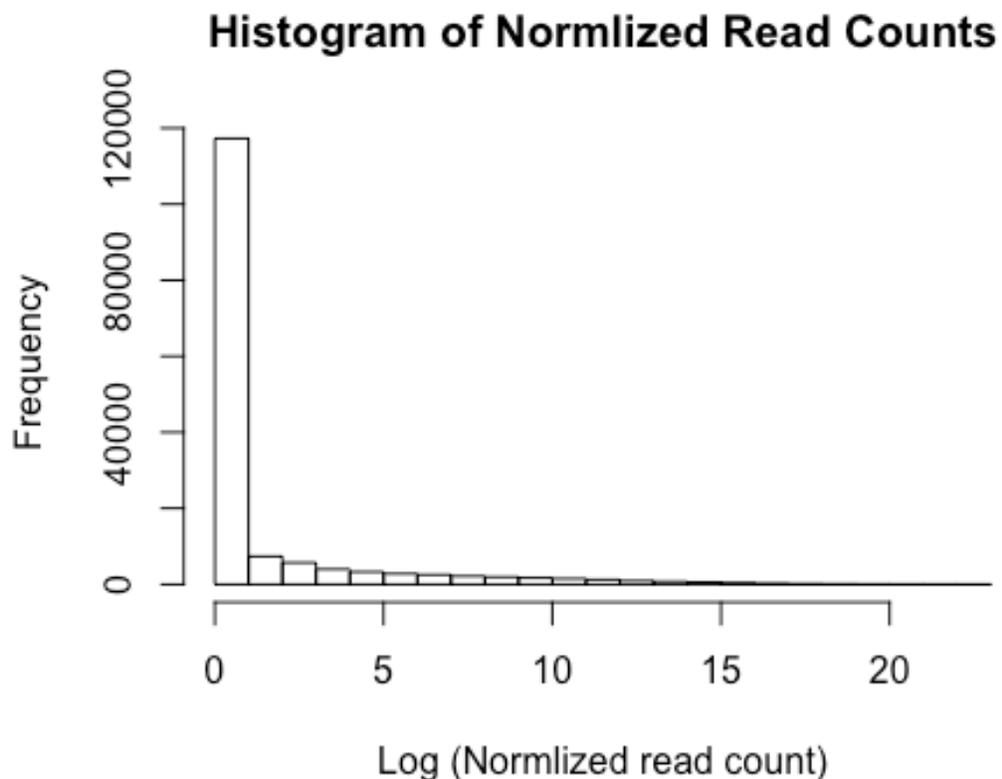
## Replace NA values with half the minimum normalized count values so that
modelling works with algorithms
(min_normalized_count<-2^min(features,na.rm = T))

## [1] 1

features[is.na(features)]<-log(min_normalized_count/2,base = 2)

## Remove low count miRNAs
hist(features,main="Histogram of Normlized Read Counts",xlab="Log (Normlized
read count)")
abline(v = -2,col='red')

```



```

nrow(features)

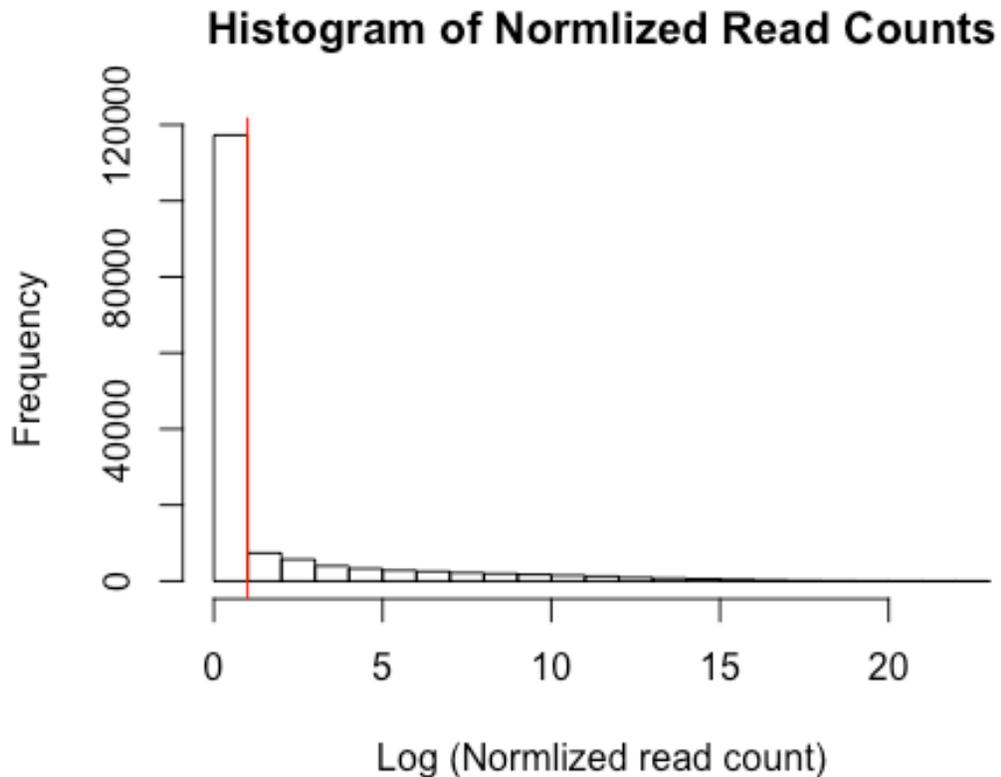
## [1] 1546

keep_features<-rowSums(features>log(min_normalized_count/2,base =
2))>=(.5*ncol(features))
sum(keep_features)

```

```
## [1] 1546

features<-features[keep_features,] ## At least half of normalized miRNA
counts should be >1.
hist(features,main="Histogram of Normlized Read Counts",xlab="Log (Normlized
read count)")
abline(v = 1,col='red')
```



```
## Write PCA plots to file
cairo_ps(file=paste0(thesis_figures_directory,FIGURE_QC_PCA.NORMAL.BYBATCH),w
idth=8,height=5.5)
plot_QC.normalized.batch$pca.p
dev.off()

## quartz_off_screen
##                2

cairo_ps(file=paste0(thesis_figures_directory,FIGURE_QC_PCA.NORMAL.BYSUBJ),wi
dth=8,height=5.5)
plot_QC.normalized.sid$pca.p
dev.off()

## quartz_off_screen
##                2
```

Prepare & Analyze the Nugent and Amsel input tables (predictors) to RF model

```
## Join the existing metdadata and predicted Amsel diagnosis
joined_meta_amselPred<-
join(pData(counts_meta.qc),data.frame(Pre_QC_ID=names(proxy_amsel_predictions
),AMSEL_prediction=proxy_amsel_predictions),by="Pre_QC_ID")

row.names(joined_meta_amselPred)<-joined_meta_amselPred$Pre_QC_ID

##Store the joined data as features_metadata
features_metadata<-ExpressionSet(features,phenoData =
AnnotatedDataFrame(joined_meta_amselPred))

## Summarize postQC counts data/metadata
summary_sid.group<-
ddply(pData(counts_meta.qc),c("SID", "BVGroup"),summarise,N=length(SID))
(summary_group.sid<-
ddply(pData(counts_meta.qc),c("BVGroup", "SID"),summarise,N=length(SID)))

##      BVGroup   SID  N
## 1      NBV   EM12  5
## 2      NBV  UAB007  5
## 3      NBV  UAB093  3
## 4      NBV  UAB102  1
## 5      NBV  UAB117  4
## 6      PBV  UAB008  3
## 7      PBV  UAB021  3
## 8      PBV  UAB022  4
## 9      PBV  UAB055  5
## 10     PBV  UAB116  4
## 11     TBV  UAB003 16
## 12     TBV  UAB005  5
## 13     TBV  UAB006 12
## 14     TBV  UAB015 14
## 15     TBV  UAB115  9
## 16     TBV  UAB121  7

(summary_group<-
ddply(pData(counts_meta.qc),c("BVGroup"),summarise,N=length(SID)))

##      BVGroup  N
## 1      NBV  18
## 2      PBV  19
## 3      TBV  63

(summary_group.uniqSID<-
ddply(unique(dplyr::select(pData(counts_meta.qc),c(SID,BVGroup))),c("BVGroup"
),summarise,N=length(SID)))
```

```

##   BVGroup N
## 1     NBV 5
## 2     PBV 5
## 3     TBV 6

(summary_group.race<-
ddply(unique(dplyr::select(pData(counts_meta.qc),c(BVGroup,Race,SID))),c("BVGroup",
"Race"),summarise,N=length(SID)))

##   BVGroup Race N
## 1     NBV     1
## 2     NBV    B 3
## 3     NBV    W 1
## 4     PBV    B 5
## 5     TBV    B 5
## 6     TBV    W 1

print(paste0("A total of ",sum(summary_sid.group$N)," samples, representing
",sum(summary_group.uniQSID$N)," unique subjects from one of 3 longitudinal
groups were used in the final analysis"))

## [1] "A total of 100 samples, representing 16 unique subjects from one of 3
longitudinal groups were used in the final analysis"

## Subset metadata for easier handling downstream
selected_metadata<-
dplyr::select(pData(features_metadata),c(SID,VAG_INT,FING_PEN,contains("SEX")
,symptoms,symptoms_nonBV,BirthControl,NUGENT_SCORE,AMSEL_prediction,Pre_QC_ID
,MENSTRUATION_NORMALIZED_PHASED,CST)) ## CSTs are *NOT* used in model input,
but are instead passed along to top mir expression plot/table downstream.
Remove CST from running in model.

##Store only counts expressed 'above 0' to reduce noise in RF models
min(exprs(features_metadata))

## [1] 0

expressed_above_0<-
exprs(features_metadata)[rowSums(exprs(features_metadata)!=0)>=(1*ncol(exprs(
features_metadata))),]

## Combine non zero log transformed normalized counts with selected metadata
as input into RF models (predictors + responses)
model_input<-data.frame(t(expressed_above_0),selected_metadata)
row.names(model_input)<-model_input$Pre_QC_ID
model_input<-dplyr::select(model_input,-Pre_QC_ID)
model_input$AMSEL_prediction<-as.character(model_input$AMSEL_prediction)
names(model_input)<-gsub(gsub(gsub(names(model_input),pattern =
"mir.",replacement = "miR-"),pattern = "hsa.",replacement = ""),replacement
="-",pattern = "\\.")

```

```

#### Subset model_input into the inputs for Amsel and Nugent RF
model_input_Amsel<-
dplyr::select(model_input[!is.na(model_input$AMSEL_prediction),],-
c(NUGENT_SCORE,CST))## Keep CSTs in model_input, just not Amsel or Nugent
input
model_input_Nugent<-
dplyr::select(model_input[!is.na(model_input$NUGENT_SCORE),],-
c(AMSEL_prediction,CST)) ## Keep CSTs in model_input, just not Amsel or
Nugent input

model_input_Amsel[is.na(model_input_Amsel)]<-0 ## RF can not have missing
values. Replacing with 0's isn't technically valid, but should have a minimal
effect on outcome
model_input_Nugent[is.na(model_input_Nugent)]<-0 ## RF can not have missing
values. Replacing with 0's isn't technically valid, but should have a minimal
effect on outcome

##ADD QC STAGE TO SRL_meta_table, then write to disk
SRL_meta_table<-left_join(SRL_meta_table,removed_samples,by="Pre_QC_ID")
print(paste0("There were 5 samples removed due to insufficient library
material or failure to sequence, and
",sum(!is.na(SRL_meta_table$QC_removal_stage) &
SRL_meta_table$QC_removal_stage %in% c("Low_Coverage","Visual Outlier")),
samples were removed due to low total miRNA reads or outliers."))

## [1] "There were 5 samples removed due to insufficient library material or
failure to sequence, and 13 samples were removed due to low total miRNA reads
or outliers."

write.csv(SRL_meta_table,file=paste0(thesis_tables_directory,TABLE_SRL_METADA
TA),row.names=F,quote=F)
SRL_seq_summary<-
ddply(SRL_meta_table,c("BVGroup","SID"),summarise,PreQC=length(QC_removal_sta
ge),PostQC=length(QC_removal_stage)-sum(as.numeric(QC_removal_stage),na.rm =
T))
SRL_seq_summary<-SRL_seq_summary[order(SRL_seq_summary$BVGroup),]
write.csv(SRL_seq_summary,file=paste0(thesis_tables_directory,TABLE_SEQSUMMAR
Y),row.names=F,quote=F)

```

Run Random Forest Models

Use RF to discover miRNAs associated w/ Lactobacillus spp. dominated communities vs CST-IV (BV-associated communities)

* Amsel- Random Forest (using predicted Amsel diagnoses as response variable)

* Nugent- Random Forest (using Nugent score as response variable)

```

## ///////////////
### Amsel-RF
## ///////////////
## See run_randomForest function above for paramaters. Note that you can load

```

a previously-built model to save time. Note that if loaded from previous, most of the input is ignored.

```
Amsel_RF<-run_randomForest(predictors_response_table = model_input_Amsel,
                             response_variable_name = "AMSEL_prediction",
                             subj_spec =T,
                             nfold = nfold,
                             nreps = npermut,
                             permute = T,
                             save_model = F,
                             file_n = "Amsel_RF",
                             load_prev_model = T,
                             verbose = F,
                             training_prop = training_prop,
                             R_script_output_directory=R_script_output_directory)

## Store top Amsel-RF features
(Amsel_RF$top_features$top_features.all<-
row.names(Amsel_RF$rfp$importance_w_pval[Amsel_RF$rfp$importance_w_pval[, "MeanDecreaseGini.pval"]<=pval_threshold |
Amsel_RF$rfp$importance_w_pval[, "MeanDecreaseAccuracy.pval"]<=pval_threshold,
]))

## [1] "let-7e" "miR-203b" "miR-184" "miR-193b" "let-7a-2" "miR-7-3"
## [7] "miR-182" "miR-183" "miR-378a" "miR-3607" "miR-100" "miR-320a"
## [13] "miR-10b" "miR-362" "miR-324" "miR-146a" "miR-16-1" "miR-4510"
## [19] "miR-500a" "miR-342"

## Calculate accuracy
NBV.accuracy.AMSEL<-100*mean(sapply(Amsel_RF$accuracy_table, function(x)
x["NBV", "NBV"])/sapply(Amsel_RF$accuracy_table, function(x) sum(x["NBV",])))
PBV.accuracy.AMSEL<-100*mean(sapply(Amsel_RF$accuracy_table, function(x)
x["PBV", "PBV"])/sapply(Amsel_RF$accuracy_table, function(x) sum(x["PBV",])))

print(paste0("The accuracy of the proxy-Amsel-RF model classification was
",round(NBV.accuracy.AMSEL,digits = 1),"% for NBV and
",round(PBV.accuracy.AMSEL,digits = 1),"% for PBV. There were
",length(Amsel_RF$top_features$top_features.all)," significant miRNAs using
proxy-Amsel RF."))

## [1] "The accuracy of the proxy-Amsel-RF model classification was 82.3% for
NBV and 80.3% for PBV. There were 20 significant miRNAs using proxy-Amsel
RF."

## ///////////
### Nugent-RF
## ///////////

Nugent_RF<-run_randomForest(predictors_response_table = model_input_Nugent,
```

```

        response_variable_name = "NUGENT_SCORE",
        subj_spec = T,
        nfold = nfolds,
        nreps = npermutes,
        permute = T,
        save_model = F,
        file_n = "Nugent_RF",
        load_prev_model = T,
        verbose = F,
        training_prop = training_prop,

R_script_output_directory=R_script_output_directory)

##Store op Nugent RF features
(Nugent_RF$top_features$top_features.all<-
row.names(Nugent_RF$rfp$importance_w_pval[Nugent_RF$rfp$importance_w_pval[, "%
IncMSE.pval"]<=pval_threshold |
Nugent_RF$rfp$importance_w_pval[, "IncNodePurity.pval"]<=pval_threshold,]))

## [1] "miR-193b" "miR-203b" "miR-324" "miR-130a" "miR-224"
## [6] "miR-182" "miR-149" "miR-3607" "miR-375" "miR-15a"
## [11] "miR-3653" "miR-21" "miR-223" "miR-15b" "miR-378a"
## [16] "miR-203a" "miR-200b" "miR-146a" "miR-205" "miR-16-2"
## [21] "miR-152" "miR-199b" "miR-20a" "miR-95" "miR-197"
## [26] "miR-365a" "miR-101-2" "miR-191" "miR-140" "miR-500a"

##Calculate accuracy/mean abs error
accuracy_table.Nugent<-Nugent_RF$accuracy_table
(mean_absolute_error.Nugent<-sum(abs(accuracy_table.Nugent$predicted-
accuracy_table.Nugent$actual))/nrow(accuracy_table.Nugent))

## [1] 1.932824

## ///////////
### Intersect Model Results
## ///////////

intersect(Nugent_RF$top_features$top_features.all,Amsel_RF$top_features$top_f
eatures.all)

## [1] "miR-193b" "miR-203b" "miR-324" "miR-182" "miR-3607" "miR-378a"
## [7] "miR-146a" "miR-500a"

print(paste0("The Nugent-RF model correctly predicted the Nugent score within
",round(mean_absolute_error.Nugent,digits = 0)," values on average. There
were ",length(Nugent_RF$top_features$top_features.all)," significant miRNAs
using Nugent-RF."))

## [1] "The Nugent-RF model correctly predicted the Nugent score within 2
values on average. There were 30 significant miRNAs using Nugent-RF."

```

```

## ///////////////
### Output Model Results
## ///////////////

### Tables
## Write model input table to file, keeping only variables that were used in
either Nugent-RF or Amsel-RF
model_input.out<-model_input[,names(model_input) %in%
unique(c(names(model_input_Amsel),names(model_input_Nugent)))]

predictor_names<-names(dplyr::select(model_input.out,-
c(AMSEL_prediction,NUGENT_SCORE,SID)))
predictor_names.miRs<-sum(grepl(predictor_names,pattern = "miR|let"))
print(paste0("The Nugent-RF and proxy-Amsel-RF models used
",length(predictor_names)," predictors including ",predictor_names.miRs,"
non-zero log2 transformed miRNA read counts and ",length(predictor_names)-
predictor_names.miRs," metadata variables as inputs to rank feature
importance (" ,TABLE_MODEL_INPUT, ")."))

## [1] "The Nugent-RF and proxy-Amsel-RF models used 178 predictors including
169 non-zero log2 transformed miRNA read counts and 9 metadata variables as
inputs to rank feature importance (TABLE_A5.csv)."

## Record which variables were used in training and testing
model_input.out[row.names(model_input.out) %in%
Nugent_RF$training_ids,"Nugent_RF_trainingTestingSet"]<-"training"
model_input.out[row.names(model_input.out) %in%
Nugent_RF$testing_ids,"Nugent_RF_trainingTestingSet"]<-"testing"

model_input.out[row.names(model_input.out) %in%
Amsel_RF$training_ids,"Amsel_RF_trainingTestingSet"]<-"training"
model_input.out[row.names(model_input.out) %in%
Amsel_RF$testing_ids,"Amsel_RF_trainingTestingSet"]<-"testing"

write.csv(model_input.out,paste0(thesis_tables_directory,TABLE_MODEL_INPUT),r
ow.names=T,quote=F)

## Combine both RF models' importance information as a single table
TABLE_RF_SUMMARY_output<-data.frame(features=names(model_input %>%
dplyr::select(-c(AMSEL_prediction,NUGENT_SCORE,SID))))
TABLE_RF_SUMMARY_output.tmp1<-
data.frame(features=row.names(Amsel_RF$rfp$importance_w_pval),AMSEL=Amsel_RF$
rfp$importance_w_pval)
TABLE_RF_SUMMARY_output.tmp2<-
data.frame(features=row.names(Nugent_RF$rfp$importance_w_pval),NUGENT=Nugent_
RF$rfp$importance_w_pval)
TABLE_RF_SUMMARY_output<-
merge(merge(TABLE_RF_SUMMARY_output.tmp1,TABLE_RF_SUMMARY_output.tmp2,all =
T),TABLE_RF_SUMMARY_output,all = T)
TABLE_RF_SUMMARY_output<-dplyr::select(TABLE_RF_SUMMARY_output,-

```

```

contains("BV"))
write.csv(TABLE_RF_SUMMARY_output, file=paste0(thesis_tables_directory, TABLE_RF_SUMMARY), row.names=F, quote=F)
write.csv(dplyr::select(Clinical_Visit_RF$importance, -
contains("BV")), file=paste0(thesis_tables_directory, TABLE_RF_SUMMARY.CV), row.names=T, quote=F)

##Store the top miRNAs for later use
top_mirs_table<-
data.frame(miRNA=union(Nugent_RF$top_features$top_features.all, Amsel_RF$top_features$top_features.all), RF_Group=0)
top_mirs_table[top_mirs_table$miRNA %in%
Nugent_RF$top_features$top_features.all, "RF_Group"]<- "Nugent-RF"
top_mirs_table[top_mirs_table$miRNA %in%
Amsel_RF$top_features$top_features.all, "RF_Group"]<- "Amsel-RF"
top_mirs_table[top_mirs_table$miRNA %in%
Amsel_RF$top_features$top_features.all & top_mirs_table$miRNA %in%
Nugent_RF$top_features$top_features.all, "RF_Group"]<- "Both"

top_mirs_table[order(top_mirs_table$RF_Group),]

##      miRNA  RF_Group
## 31  let-7e  Amsel-RF
## 32  miR-184  Amsel-RF
## 33  let-7a-2  Amsel-RF
## 34  miR-7-3  Amsel-RF
## 35  miR-183  Amsel-RF
## 36  miR-100  Amsel-RF
## 37  miR-320a  Amsel-RF
## 38  miR-10b  Amsel-RF
## 39  miR-362  Amsel-RF
## 40  miR-16-1  Amsel-RF
## 41  miR-4510  Amsel-RF
## 42  miR-342  Amsel-RF
## 1   miR-193b   Both
## 2   miR-203b   Both
## 3   miR-324    Both
## 6   miR-182    Both
## 8   miR-3607   Both
## 15  miR-378a   Both
## 18  miR-146a   Both
## 30  miR-500a   Both
## 4   miR-130a  Nugent-RF
## 5   miR-224  Nugent-RF
## 7   miR-149  Nugent-RF
## 9   miR-375  Nugent-RF
## 10  miR-15a  Nugent-RF
## 11  miR-3653  Nugent-RF
## 12  miR-21   Nugent-RF
## 13  miR-223  Nugent-RF

```

```

## 14 miR-15b Nugent-RF
## 16 miR-203a Nugent-RF
## 17 miR-200b Nugent-RF
## 19 miR-205 Nugent-RF
## 20 miR-16-2 Nugent-RF
## 21 miR-152 Nugent-RF
## 22 miR-199b Nugent-RF
## 23 miR-20a Nugent-RF
## 24 miR-95 Nugent-RF
## 25 miR-197 Nugent-RF
## 26 miR-365a Nugent-RF
## 27 miR-101-2 Nugent-RF
## 28 miR-191 Nugent-RF
## 29 miR-140 Nugent-RF

## Text
print(paste0("A total of
",length(top_mirs_table[top_mirs_table$RF_Group=="Both","miRNA"]), " miRNAs
were common to both proxy-Amsel-BV and Nugent-BV Random Forest models: ",
str_c(top_mirs_table[top_mirs_table$RF_Group=="Both","miRNA"],collapse="
"),"."))

## [1] "A total of 8 miRNAs were common to both proxy-Amsel-BV and Nugent-BV
Random Forest models: miR-193b, miR-203b, miR-324, miR-182, miR-3607, miR-
378a, miR-146a, miR-500a."

print(paste0("There were
",nrow(top_mirs_table[top_mirs_table$RF_Group=="Amsel-RF",]), " and
",nrow(top_mirs_table[top_mirs_table$RF_Group=="Nugent-RF",]), " statistically
significant miRNAs unique to each of proxy-Amsel classification (proxy-Amsel-
RF) and Nugent score regression (Nugent-RF), respectively"))

## [1] "There were 12 and 22 statistically significant miRNAs unique to each
of proxy-Amsel classification (proxy-Amsel-RF) and Nugent score regression
(Nugent-RF), respectively"

### Figures
dplyr::select(data.frame(Amsel_RF$rfp$importance_w_pval[rownames(Amsel_RF$rfp
p$importance_w_pval) %in% top_mirs_table[top_mirs_table$RF_Group %in%
c("Amsel-RF","Both"),"miRNA"],]),contains("Mean"))

##           MeanDecreaseAccuracy MeanDecreaseAccuracy.pval MeanDecreaseGini
## let-7e           5.607872           0.002195609           0.9470099
## miR-203b          4.630648           0.003193613           0.7717181
## miR-184           4.224487           0.003792415           0.5855550
## miR-193b          3.882583           0.005588822           0.4847252
## let-7a-2          3.626207           0.006786427           0.3909026
## miR-7-3           3.378600           0.009181637           0.4143229
## miR-182           3.235782           0.010778443           0.4662138
## miR-183           3.065067           0.012375250           0.4495199
## miR-378a          2.909091           0.012375250           0.3701688

```

```

## miR-3607          2.801747          0.016367265          0.3678346
## miR-100           2.687550          0.020958084          0.3507171
## miR-320a         2.597080          0.022754491          0.2273535
## miR-10b          2.506643          0.026746507          0.3665795
## miR-362          2.426444          0.026746507          0.2311134
## miR-324          2.327208          0.034131737          0.2919702
## miR-146a         2.257027          0.038123752          0.3095903
## miR-16-1         2.154749          0.043313373          0.2736763
## miR-4510         2.096317          0.046706587          0.2717785
## miR-500a         2.020440          0.048103792          0.2481224
## miR-342          1.973390          0.049900200          0.1733185
##                MeanDecreaseGini.pval
## let-7e           0.002195609
## miR-203b         0.002994012
## miR-184          0.006986028
## miR-193b         0.020558882
## let-7a-2         0.057285429
## miR-7-3          0.041916168
## miR-182          0.022355289
## miR-183          0.069261477
## miR-378a         0.047704591
## miR-3607         0.102395210
## miR-100          0.135728543
## miR-320a         0.257285429
## miR-10b          0.103393214
## miR-362          0.250299401
## miR-324          0.258283433
## miR-146a         0.185828343
## miR-16-1         0.150698603
## miR-4510         0.167664671
## miR-500a         0.202994012
## miR-342          0.415369261

```

```

Amsel_RF.plot<-
plot_importance(dplyr::select(data.frame(Amsel_RF$rfp$importance_w_pval[row.n
ames(Amsel_RF$rfp$importance_w_pval) %in%
top_mirs_table[top_mirs_table$RF_Group %in% c("Amsel-
RF","Both"),"miRNA"],)),contains("Accuracy"),ntopfeats = 20,rankBy =
"MeanDecreaseAccuracy",nfeats = 20,model_name = "proxy-Amsel",size_font = 20)

```

```

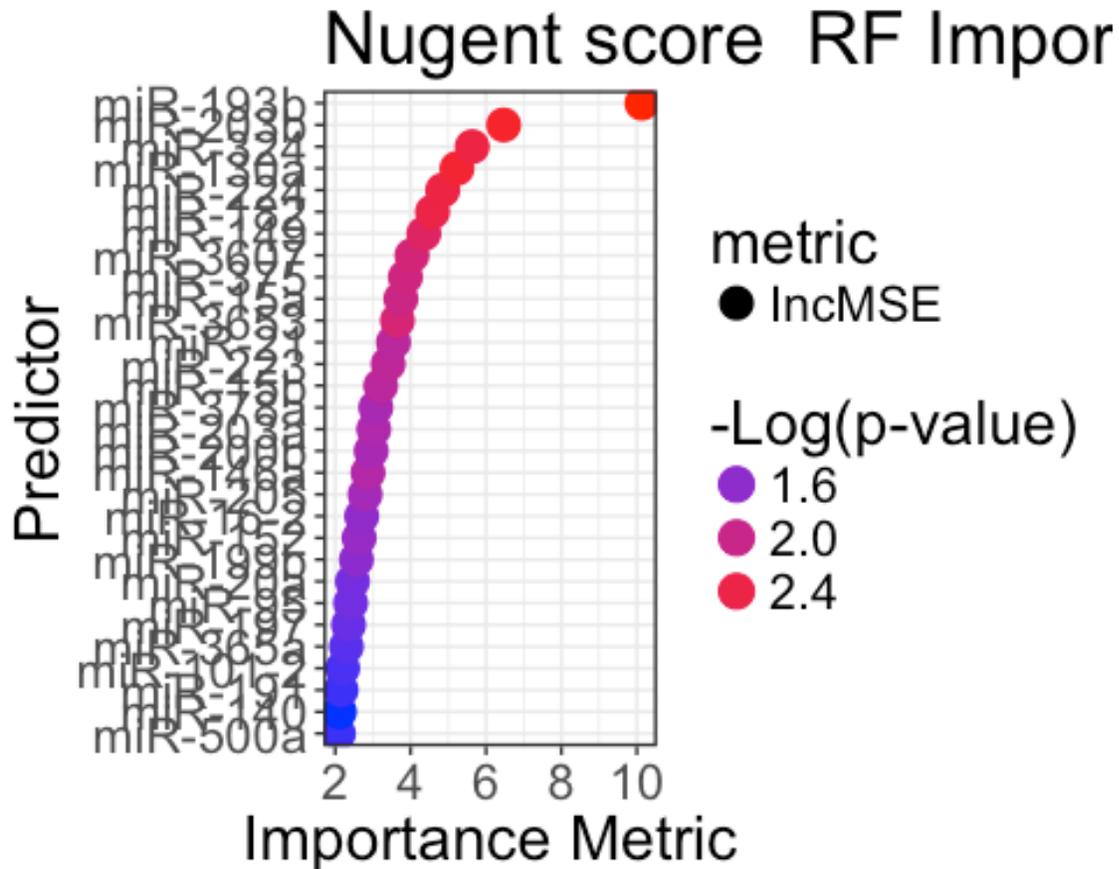
Nugent_RF.plot<-
plot_importance(dplyr::select(data.frame(Nugent_RF$rfp$importance_w_pval[row.
names(Nugent_RF$rfp$importance_w_pval) %in%
top_mirs_table[top_mirs_table$RF_Group %in% c("Nugent-
RF","Both"),"miRNA"],)),contains("MSE"),ntopfeats = 30,rankBy =
"IncMSE",nfeats = 30,model_name = "Nugent score",size_font = 20)

```

```

plot(Amsel_RF.plot)

```

```

cairo_ps(file=paste0(thesis_figures_directory,FIGURE_RF_IMPORTANCE_AMSEL),
width=11, height=8.5)
plot(Amsel_RF.plot)
dev.off()

## quartz_off_screen
##                2

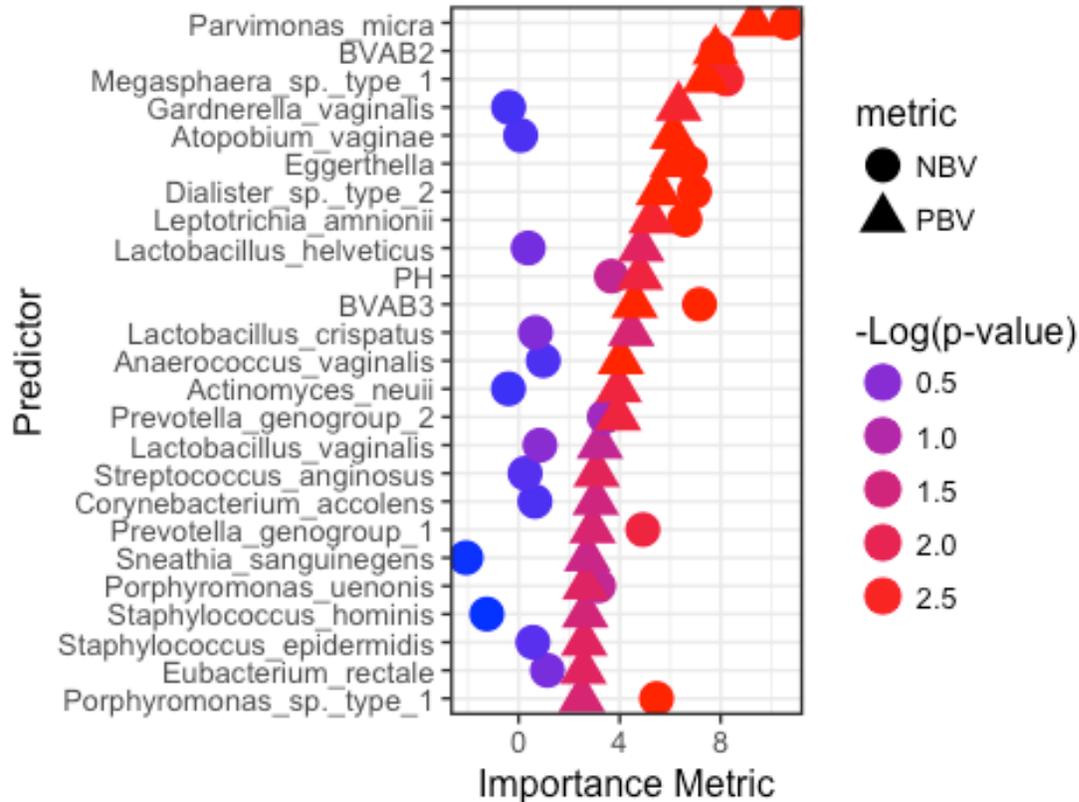
cairo_ps(file=paste0(thesis_figures_directory,FIGURE_RF_IMPORTANCE_NUGENT),
width=11, height=8.5)
plot(Nugent_RF.plot)
dev.off()

## quartz_off_screen
##                2

plot_importance(dplyr::select(Clinical_Visit_RF$importance,contains("BV")),nt
opfeats = 20,rankBy = "PBV",model_name = "Amsel")

```

Amsel RF Importance Plot for the



```
## Plot Expression of each RF mir as a function of Nugent score & proxy-Amsel prediction
```

```
topmir_plot_data<-
model_input[!is.na(model_input$NUGENT_SCORE),names(model_input) %in%
c("NUGENT_SCORE","AMSEL_prediction","CST",as.character(top_mirs_table$miRNA[top_mirs_table$RF_Group %in% "Both"]))]
topmir_plot_data$CST<-as.character(topmir_plot_data$CST)
```

```
topmir_plot_data<-melt(topmir_plot_data,id.vars =
c("NUGENT_SCORE","AMSEL_prediction","CST"))
```

```
topmir_plot_data$variable<-gsub(topmir_plot_data$variable,pattern =
"hsa.mir.",replacement = "miR-")
```

```
## Calcualte linear model fit for miRNA expression data vs Nugent score to rank miR expression plots in figure
```

```
fitness<-apply(model_input[,names(model_input) %in%
unique(topmir_plot_data$variable)],MARGIN = 2,function(s)
summary(lm(model_input$NUGENT_SCORE~s)))
```

```

fitness.r2<-data.frame(adjr2=unlist(lapply(fitness,function(x)
round(x$adj.r.squared,digits = 3))))
fitness.r2$variable<-row.names(fitness.r2)

print(fitness.r2[order(abs(fitness.r2$adjr2),decreasing = T),])

##          adjr2 variable
## miR-193b 0.379 miR-193b
## miR-182 0.214 miR-182
## miR-203b 0.182 miR-203b
## miR-378a 0.140 miR-378a
## miR-3607 0.134 miR-3607
## miR-324 0.124 miR-324
## miR-500a 0.075 miR-500a
## miR-146a 0.001 miR-146a

topmir_plot_data<-join(topmir_plot_data,fitness.r2)

## Joining by: variable

topmir_plot_data$variable<-factor(x=topmir_plot_data$variable,levels =
unique(topmir_plot_data[order(abs(topmir_plot_data$adjr2),decreasing =
T),c("variable")]),ordered = T)

## Liner correlation coeff for miRNA expression data:
paste(apply(unique(dplyr::select(topmir_plot_data,c(variable,adjr2))),MARGIN
= 1,FUN = function(x) paste(sep = ":" ,x[1],x[2])),sep=",")

## [1] "miR-146a:0.001" "miR-182:0.214" "miR-193b:0.379" "miR-203b:0.182"
## [5] "miR-324:0.124" "miR-3607:0.134" "miR-378a:0.140" "miR-500a:0.075"

## Collapse CSTs to a single CST super type
topmir_plot_data[topmir_plot_data$CST %in% c("III-A","III-B"),"CST"]<-"III"
topmir_plot_data[topmir_plot_data$CST %in% c("I-A","I-B"),"CST"]<-"I"

## Plot the miRNA expression figure
PLOT_EXPRESSION<-
  ggplot(topmir_plot_data,aes(x=NUGENT_SCORE,y=value))+
  stat_smooth(method = "lm",col='#FF5733',se=F)+ #
  geom_point(aes(col=as.factor(CST),pch=AMSEL_prediction),size=3,show.legend
= T)+
  facet_wrap(~variable,scales = "free_y",nrow=2)+
  ylab("Normalized Expression")+xlab("Nugent Score")+
  mBio+
  theme(text = element_text(size=20),axis.text.x = element_text(size=15))+
  scale_x_continuous(breaks=0:10)+
  scale_color_manual(values = cst.colors,guide = guide_legend(title =
"CST"))+
  guides(pch=guide_legend(title="proxy Amsel Prediction"))

cairo_ps(paste0(thesis_figures_directory,FIGURE_TOPMIRS),width = 16,height =

```

```

8) #
plot(PLOT_EXPRESSION)
dev.off()

## quartz_off_screen
##                2

library(Boruta)

## Loading required package: ranger

##
## Attaching package: 'ranger'

## The following object is masked from 'package:randomForest':
##
##      importance

postscript(paste0(R_script_output_directory,"Boruta_optimization.ps"))
Nugent_RF.Boruta<-Boruta(x=dplyr::select(model_input_Nugent,-
c(SID,NUGENT_SCORE)),y=model_input_Nugent$NUGENT_SCORE)
plot(Nugent_RF.Boruta,main="Nugent-RF Boruta unimportant feature removal")
Nugent_RF.Boruta$finalDecision[Nugent_RF.Boruta$finalDecision=="Confirmed"]

## miR-101-1 miR-101-2 miR-128-1 miR-128-2 miR-130a miR-142
## Confirmed Confirmed Confirmed Confirmed Confirmed Confirmed
## miR-143 miR-149 miR-15a miR-182 miR-193b miR-194-2
## Confirmed Confirmed Confirmed Confirmed Confirmed Confirmed
## miR-199a-1 miR-199b miR-203a miR-203b miR-205 miR-223
## Confirmed Confirmed Confirmed Confirmed Confirmed Confirmed
## miR-320b-1 miR-324 miR-3607 miR-365a miR-378a miR-486
## Confirmed Confirmed Confirmed Confirmed Confirmed Confirmed
## Levels: Tentative Confirmed Rejected

Amsel_RF.Boruta<-Boruta(x=dplyr::select(model_input_Amsel,-
c(SID,AMSEL_prediction)),y=as.factor(model_input_Amsel$AMSEL_prediction))
plot(Amsel_RF.Boruta,main="proxy-Amsel-RF Boruta unimportant feature
removal")
Amsel_RF.Boruta$finalDecision[Amsel_RF.Boruta$finalDecision=="Confirmed"]

## let-7a-2 miR-100 miR-146a miR-182 miR-183 miR-184 miR-193b
## Confirmed Confirmed Confirmed Confirmed Confirmed Confirmed Confirmed
## miR-203b miR-223 miR-320a miR-3607 miR-486
## Confirmed Confirmed Confirmed Confirmed Confirmed
## Levels: Tentative Confirmed Rejected

intersect(names(Nugent_RF.Boruta$finalDecision[Nugent_RF.Boruta$finalDecision
=="Confirmed"]),names(Amsel_RF.Boruta$finalDecision[Amsel_RF.Boruta$finalDeci
sion=="Confirmed"]))

## [1] "miR-182" "miR-193b" "miR-203b" "miR-223" "miR-3607" "miR-486"

```

```
dev.off()
```

```
## quartz_off_screen  
## 2
```

Map miRNAs to Gene Targets and GO Process (miR-GO-Target) (Table 2)

Using experimentally validated miRNA targets, map to direct GO processes to discover common & unique functions of top miRNAs. 1. Output gene target list for each top miRNA. 2. Count the frequency of each target for each GO process & miRNA. Plot.

```
### miRNA <-> Target (gene) <-> Gene Ontology Process (direct)  
### miRNA <- - - - - - - - > Gene Ontology Process (direct)  
  
## Read in miRTarBase Experimentally Validated Target List  
mir_targets<-  
read.table(paste0(R_script_input_directory,"miRTarBase_SE_WR_homosapiens.txt"),  
header = T,sep="\t",stringsAsFactors = F) %>% filter(Species..miRNA=="Homo  
sapiens")  
nrow(mir_targets)  
  
## [1] 7310  
  
## Format targets list,pattern = "hsa.",replacement = "")  
mir_targets$miRNA.format<-gsub(gsub(gsub(mir_targets$miRNA,pattern =  
"mir.",replacement = "miR-"),replacement = "-"),pattern = "\\."),pattern = "-  
.p",replacement = "-")  
  
## Map miRNA names to target List  
for(mir in as.character(top_mirs_table$miRNA)){  
  target_list<-mir_targets[grep(mir_targets$miRNA.format,pattern =  
paste0("hsa-",mir,"-"),ignore.case = T,perl =  
T),c("miRNA","Target.Gene","References..PMID.")]  
  #if(!nrow(target_list)==0){target_list$source<-"canonical"}  
  if(!nrow(target_list)==0 ){  
    top_mirs_table[top_mirs_table$miRNA==mir,"targets"]<-  
str_c(sort(unique(target_list$Target.Gene)),collapse = ",")  
    top_mirs_table[top_mirs_table$miRNA==mir,"sanityCheck"]<-  
str_c(sort(unique(target_list$miRNA)),collapse = ",")  
    top_mirs_table[top_mirs_table$miRNA==mir,"PMID"]<-  
str_c(sort(unique(target_list$References..PMID.)),collapse = ",")  
  }  
}  
  
write.table(top_mirs_table,file=paste0(thesis_tables_directory,TABLE_MIR_TARG  
ETS),sep="\t",quote = F,row.names = F) ## Needs to be tab delimited due to  
"," separating gene targets.
```

```

### Read in miR-Go-Gene mapping table mapping targets/miRNAs to GO processes.
Gene targets mapped to miRNAs using a Python script.
mir_go<-
read.csv(paste0(R_script_input_directory,"miR_GODirect_Gene_Map.csv"),strings
AsFactors = F)

## Filter just the miR-GO map to 'top' miRNAs discovered above belonging to
both RF model results
mir_go<-mir_go[mir_go$mir %in%
as.character(filter(top_mirs_table,RF_Group=="Both")$miRNA),]
ddply(unique(dplyr::select(mir_go,c(mir,gene))),c("mir"),summarise,n=length(g
ene))

##      mir  n
## 1 miR-146a 48
## 2 miR-182 46
## 3 miR-193b 12
## 4 miR-324  6
## 5 miR-378a 16
## 6 miR-500a  3

## Read in miR-target counts, created in outside script.
mir_targetCnt<-
read.csv(paste0(R_script_input_directory,"miR_Target_Counts.csv"),stringsAsFa
ctors = F)
mir_targetCnt[mir_targetCnt$mir %in%
top_mirs_table[top_mirs_table$RF_Group=="Both","miRNA"],]

##      mir number_targets
## 9 miR-378a             16
## 13 miR-193b            13
## 19 miR-203b             1
## 21 miR-3607             1
## 22 miR-324              6
## 27 miR-182             48
## 33 miR-146a            51
## 41 miR-500a             3

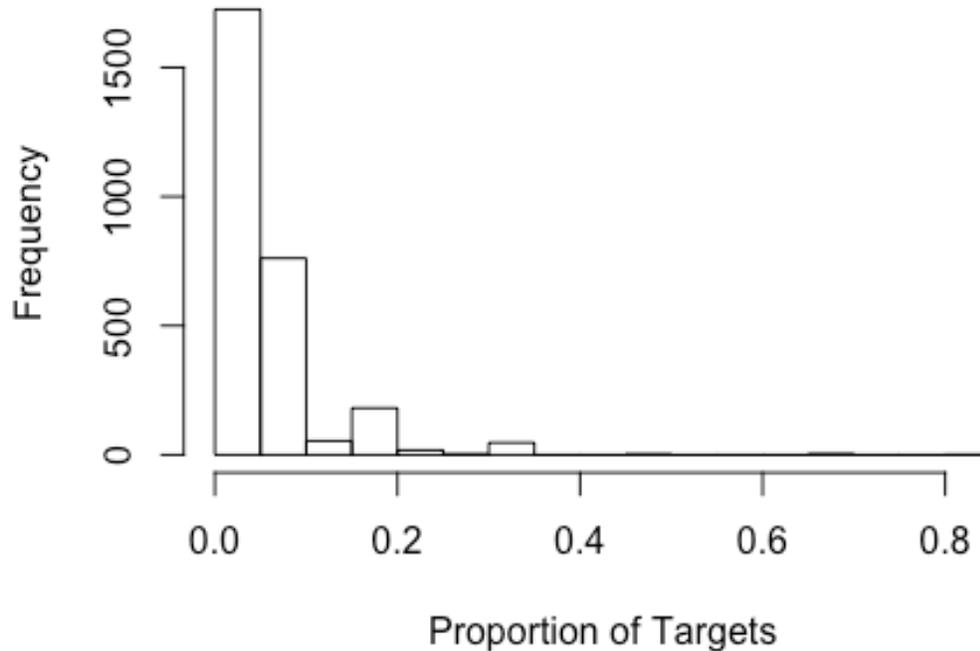
## Determine the number of genes associated with each miRNA- need to
determine how many GO processes to plot in figure
mir_go.counts<-data.frame(table(dplyr::select(mir_go,c(mir,GO))))
mir_go.miR193b<-mir_go[mir_go$mir=="miR-193b",]

mir_go.counts<-merge(mir_go.counts,mir_targetCnt)
mir_go.counts$Freq_Targets<-mir_go.counts$Freq/mir_go.counts$number_targets

hist(mir_go.counts$Freq_Targets[!mir_go.counts$Freq_Targets==0],main="Histogram
of Non-zero miRNA Target Frequency",xlab="Proportion of Targets")

```

Histogram of Non-zero miRNA Target Frequency



```
## Somewhere above 0.2 might capture the interesting/common GO processes
## across miRNAs.

## Make miR-GO counts into wide format (i.e., each GO process is a column,
## miRNAs are rows, values are proportion of targets in that GO/miRNA
## combination)
mir_go.wide<-tidyr::spread(dplyr::select(mir_go.counts,-
c(Freq,number_targets)),key = GO,value = Freq_Targets)

row.names(mir_go.wide)<-mir_go.wide$mir
mir_go.wide<-dplyr::select(mir_go.wide,-mir)

## Get a better idea of how to plot common GO processes by determining
## cutoffs for target proportions
## Balance between proportion of targets across all miRNAs and total #
## across targets

## create dummy dataframe- i and j hold the tuning parameters
df1<-
data.frame(i=rep(seq(0,max(mir_go.wide),max(mir_go.wide)/10),each=10),j=rep(s
eq(0,1,.1),times=10))

## Calculate number of columns (nc) for i and j cutoffs
```

```

for(a in 1:nrow(df1)){
  i<-df1[a,"i"]
  j<-df1[a,"j"]
  ## number of columns w/ propotion >i and of those, number greater than j of
total miRNAs
  nc<-ncol(mir_go.wide[,colSums(mir_go.wide>i)>=j*nrow(mir_go.wide)])
  if(is.null(nc)){nc<-0}
  df1[a,"ncol"]<-nc
}
(optimal_cutoffs<-df1[df1$ncol==max(df1[df1$ncol<=25 &
df1$ncol>=5,"ncol"]),]) ## What are the optimal cutoffs for somewhere between
5 and 25 GO processes on the plot?

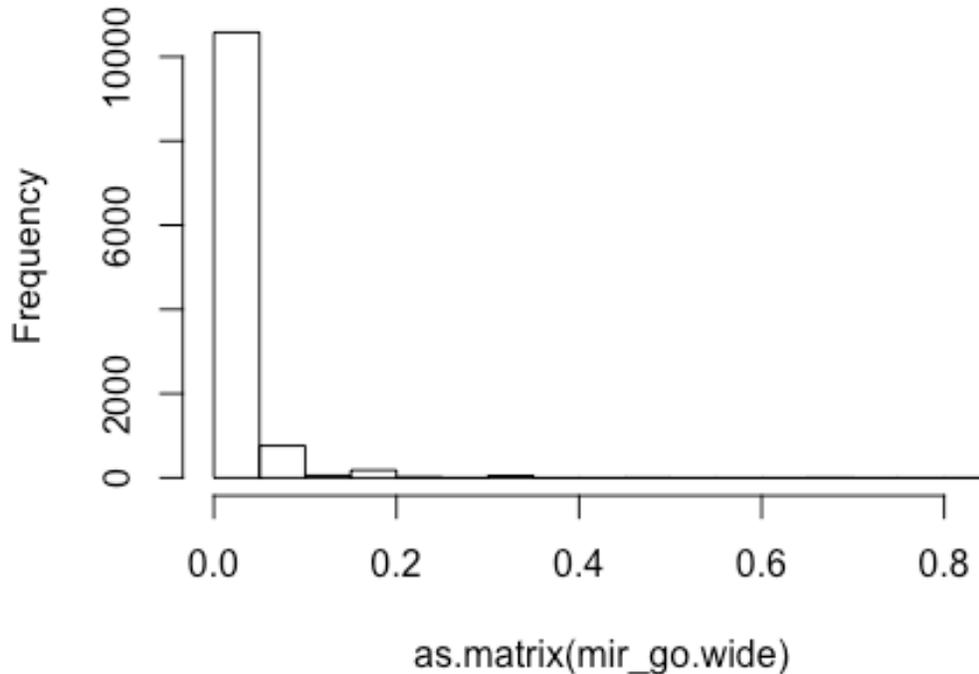
##   i   j ncol
## 8 0 0.7   19
## 9 0 0.8   19

## Take the max i/j pair as the cutoffs:
i<-0
j<-0.8
## Note that 19 in the "nc" column is the number of GO processes that will be
plotted

## Therefore, the proportion of genes should be >0 across at least 0.8*6 =4.8
miRNAs.
hist(as.matrix(mir_go.wide))

```

Histogram of as.matrix(mir_go.wide)



```
## Apply optimal cutoffs to miR-GO (wide) data
mir_go.wide.filtered<-
mir_go.wide[,colSums(mir_go.wide>i)>=j*nrow(mir_go.wide)]

mir_go.wide.mir193b<-mir_go.wide[row.names(mir_go.wide)=="miR-193b",]

## The names of the GO processes to be used in plot:
names(mir_go.wide.filtered[,order(colSums(mir_go.wide.filtered>0),decreasing
= T)])

## [1] "GO:000122~negative regulation of transcription from RNA polymerase
II promoter"
## [2] "GO:0001666~response to hypoxia"
## [3] "GO:0001701~in utero embryonic development"
## [4] "GO:0001934~positive regulation of protein phosphorylation"
## [5] "GO:0006351~transcription<COMMA> DNA-templated"
## [6] "GO:0006355~regulation of transcription<COMMA> DNA-templated"
## [7] "GO:0006366~transcription from RNA polymerase II promoter"
## [8] "GO:0006955~immune response"
## [9] "GO:0007049~cell cycle"
## [10] "GO:0007165~signal transduction"
## [11] "GO:0007275~multicellular organism development"
## [12] "GO:0008284~positive regulation of cell proliferation"
```

```

## [13] "GO:0010628~positive regulation of gene expression"
## [14] "GO:0030335~positive regulation of cell migration"
## [15] "GO:0043066~negative regulation of apoptotic process"
## [16] "GO:0043547~positive regulation of GTPase activity"
## [17] "GO:0045893~positive regulation of transcription<COMMA> DNA-
templated"
## [18] "GO:0045944~positive regulation of transcription from RNA polymerase
II promoter"
## [19] "GO:0051091~positive regulation of sequence-specific DNA binding
transcription factor activity"

mir_go.wide.filtered$mir<-row.names(mir_go.wide.filtered)
mir_go.wide.filtered.m<-melt(mir_go.wide.filtered,id.vars = c("mir"))
mir_go.wide.filtered.m$GO_only<-gsub(mir_go.wide.filtered.m$variable,pattern
= "~.*",replacement = "") ## Store the GO number only in case needed

## For this plot, summarize the mean gene target proportion and # of miRNAs
for each GO process. Use this to rank GO processes in plot
(mir_go.wide.filtered.summary<-
ddply(mir_go.wide.filtered.m,c("variable"),summarise,s=sum(value),n=sum(value
>0)))

##
variable
## 1          GO:0000122~negative regulation of transcription from RNA
polymerase II promoter
## 2
GO:0001666~response to hypoxia
## 3          GO:0001701~in utero
embryonic development
## 4          GO:0001934~positive regulation of
protein phosphorylation
## 5
GO:0006351~transcription<COMMA> DNA-templated
## 6          GO:0006355~regulation of
transcription<COMMA> DNA-templated
## 7          GO:0006366~transcription from RNA
polymerase II promoter
## 8
GO:0006955~immune response
## 9
GO:0007049~cell cycle
## 10
GO:0007165~signal transduction
## 11
GO:0007275~multicellular organism development
## 12          GO:0008284~positive regulation
of cell proliferation
## 13          GO:0010628~positive
regulation of gene expression

```

```

## 14                                GO:0030335~positive
regulation of cell migration
## 15                                GO:0043066~negative
regulation of apoptotic process
## 16                                GO:0043547~positive
regulation of GTPase activity
## 17                                GO:0045893~positive regulation of
transcription<COMMA> DNA-templated
## 18                                GO:0045944~positive regulation of transcription from RNA
polymerase II promoter
## 19 GO:0051091~positive regulation of sequence-specific DNA binding
transcription factor activity
##          s n
## 1  1.5555241 6
## 2  0.7194570 5
## 3  0.6161388 5
## 4  0.5043363 5
## 5  0.8864065 5
## 6  0.7283183 5
## 7  0.7479261 5
## 8  0.7077677 5
## 9  0.6761878 5
## 10 0.9829374 5
## 11 0.5131976 5
## 12 0.9694570 5
## 13 0.6918363 5
## 14 0.6281109 5
## 15 1.0101810 5
## 16 0.4457956 5
## 17 1.3226810 5
## 18 1.4440045 5
## 19 0.3881976 5

## T0 make plot simpler to read, group GO terms in logical super groupings.
This is done using a map between GO group and GO term, created outside of
script
group_go_map<-
read.csv(file=paste0(R_script_input_directory,"GO_Process_Grouping_Map.csv"))

## How many GO terms in each super group
(group_freq<-data.frame(table(group_go_map$Group)))

##          Var1 Freq
## 1    Cell cycle    4
## 2  Development    2
## 3    Hypoxia      1
## 4    Immune       1
## 5    Signaling    3
## 6 Transcription    8

```

```

names(group_freq)<-c("Group","n_group")
# MErge the super groups and data found above
group_go_map<-merge(group_go_map,group_freq)

## Order by grouping, then by mean target proportion within each group
group_go_map<-group_go_map[order(group_go_map$n_group,
group_go_map$s,decreasing = T),]

#FOrmat the GO term for plot
mir_go.wide.filtered.m$GO_format<-
factor(mir_go.wide.filtered.m$variable,levels =group_go_map$variable ,ordered
= T,labels = group_go_map$Go_format)

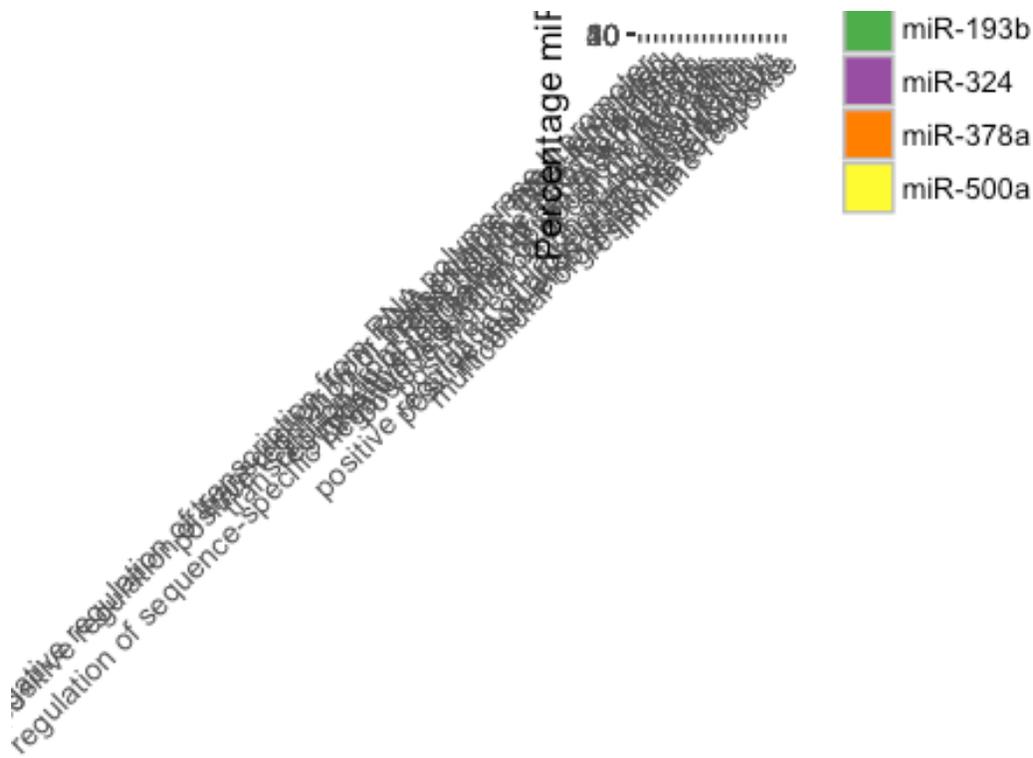
PLOT_MIR_TARGET_GO_PROP<-
ggplot(mir_go.wide.filtered.m,aes(x=GO_format,y=100*value,fill=mir))+geom_bar
(width=.7,position = "dodge",col="grey",stat="identity")+
  mBio+
  theme(axis.text.x = element_text(angle=45,hjust =
1,size=10),text=element_text(size=12),plot.margin =
unit(c(10,25,10,170),units = "pt"))+
  scale_fill_brewer(type = "qual",palette = "Set1")+
  xlab("GO Term")+
  ylab("Percentage miRNA Targets")

cairo_ps(paste0(thesis_figures_directory,FIGURE_MIR_TARGETS_GO),width =
11,height = 8.5)
plot(PLOT_MIR_TARGET_GO_PROP)
dev.off()

## quartz_off_screen
##                2

plot(PLOT_MIR_TARGET_GO_PROP)

```



GO Term

```
## NOTE: the # of miR targets table is added later. See above for table.

## Output miR-Target-GO table for miR-193b only
mir_go.wide.filtered.m.193b<-filter(mir_go.wide.filtered.m,mir=="miR-193b")

write.table(dplyr::select(mir_go.wide.filtered.m.193b,c(mir,GO_format,value))
,file =
paste0(thesis_tables_directory,TABLE_TOPMIRS),sep="\t",quote=F,row.names = F)

## Make a GO process bar chart for miR-193b only (for thesis presentation,
not used in thesis document)
mir193b_GOcount_groups<-
data.frame(table(dplyr::select(unique(dplyr::select(merge(mir_go.miR193b,grou
p_go_map,by.x="GO",by.y="variable"),c(Group,gene))),Group)))
mir193b_GOcount_groups$Var1<-
factor(mir193b_GOcount_groups$Var1,levels=mir193b_GOcount_groups[order(mir193
b_GOcount_groups$Freq,decreasing = T),"Var1"],ordered = T)

mir193b_GOcount_groups.plot<-
ggplot(mir193b_GOcount_groups,aes(x=Var1,y=Freq))+geom_bar(width=.7,position
= "dodge",col="grey",stat="identity")+
mBio+
theme(axis.text.x = element_text(angle=45,hjust =
```

```

1),text=element_text(size=25),plot.margin = unit(c(10,25,10,170),units =
"pt"))+
  scale_fill_brewer(type = "qual",palette = "Set1")+
  scale_y_continuous(breaks=c(2,4,6,8))+
  xlab("GO process")+
  ylab("Number miR-193b targets")

cairo_ps(paste0(R_script_output_directory,"miR193b_GOProcess_presentation.eps
"),width = 11,height = 8.5)
plot(mir193b_GOcount_groups.plot)
dev.off()

## quartz_off_screen
##           2

```

Function of miR-193b and in vitro Experimentation, Implications

Perform further experimentation on miR-193b, one of the top miRNAs found in the discovery phase above.

Validate SmallRNASeq Results using qPCR

::select 5 swabs each from NBV and PBV subjects, then measure relative hsa-mir-193b expression

```

## Loads qPCR Data from 5 NBV and PBV subjects to Validate miRNA-Seq
load(paste0(R_script_input_directory,"miRNA_qPCR_Validation.RData"))
miR_qPCR_results.deltaCt<-
miR_qPCR_results.deltaCt[!is.nan(miR_qPCR_results.deltaCt$deltaCt),] ##
remove NA from data

## Create a mixed-effects linear model as a function of BVGroup of dCt
values, using subject ID as the random effect
seq_validation.lme <- lme(deltaCt ~ BVGroup,
                        random = ~ 1|SID,
                        data = miR_qPCR_results.deltaCt)

## Coefficient of PBV relative to NBV in model + p value
effect_PBV<-seq_validation.lme$coefficients$fixed["BVGroupPBV"]
dct.summary<-
ddply(miR_qPCR_results.deltaCt,c("BVGroup"),summarise,mean=mean(deltaCt))
2^(dct.summary$mean[2]-dct.summary$mean[1])

## [1] 3.019121

summary(seq_validation.lme)

## Linear mixed-effects model fit by REML
## Data: miR_qPCR_results.deltaCt

```

```

##           AIC           BIC      logLik
##  19.19847 18.36551 -5.599237
##
## Random effects:
## Formula: ~1 | SID
##           (Intercept) Residual
## StdDev:    0.6267906 0.107014
##
## Fixed effects: deltaCt ~ BVGroup
##           Value Std.Error DF   t-value p-value
## (Intercept) -1.682609 0.3179302  5 -5.292385  0.0032
## BVGroupPBV   1.432459 0.4849846  5  2.953617  0.0318
## Correlation:
##           (Intr)
## BVGroupPBV -0.656
##
## Standardized Within-Group Residuals:
##           Min           Q1           Med           Q3           Max
## -0.62114365 -0.18105809  0.03265107  0.08253511  0.79729434
##
## Number of Observations: 8
## Number of Groups: 7

```

miR-193b qPCR Time-Course in VK2 monolayer after BCS Exposure

qPCR is measured with QIAGEN miScript II kit, BCSs are:

- * L. crispatus
- * L. jensenii
- * L. iners
- * G. vaginalis
- * Media control

20% BCSs exposed to VK2 monolayer epithelial cells for:

- * 0.5 hours
- * 1 hour
- * 4 hours
- * 11 hours
- * 13 hours
- * 22 hours

Primers:

- * hsa-miR-193b-3p
- * hsa-miR-RNU6 (normalization control)

Calculate mean & s.d. Ct values for each BCS/time point.

```

load(file=paste0(R_script_input_directory,"qPCR_time_course.Rdata")) ## Ct
values
load(file=paste0(R_script_input_directory,"qPCR_sigtest.Rdata")) ## Container
for holding sig testing comparisons/results

```

```

qPCR_time_course.summary<-
ddply(qPCR_time_course,c("Study","ExposureTime","BCS"),summarise,m_193b=mean(
miR193b_Ct,na.rm = T),sd_193b=sd(miR193b_Ct,na.rm = T),
                                m_R6=mean(RNU6_Ct,na.rm =
T),sd_R6=sd(RNU6_Ct,na.rm = T))

## Calcualte delta Ct using RNU6 as endogenous control. Calcualte standard
deviation using propagation of error
qPCR_time_course.summary$deltaCt.RNU6<-qPCR_time_course.summary$m_193b-
qPCR_time_course.summary$m_R6
qPCR_time_course.summary$deltaCt.RNU6.sd<-
sqrt(qPCR_time_course.summary$sd_193b^2+qPCR_time_course.summary$sd_R6^2)
#

## calcualte signifigance using delta Ct mean + s.d., which uses a modified
t.test2 function.
## Loop through the comparisons and test pair-wise
for(sigtest in 1:nrow(qpcr_sigtest)){

  timei<-as.character(qpcr_sigtest[sigtest,"ExposureTime"]) # Time point
  x1<-as.character(qpcr_sigtest[sigtest,"comparison"]) ## Comparison sample
  y1<-as.character(qpcr_sigtest[sigtest,"reference"]) ## Reference sample

  ## When comparing samples, make sure to use the correct study associated
  with BCS
  study.x<-as.character(qpcr_sigtest[sigtest,"Study.comp"])
  study.y<-as.character(qpcr_sigtest[sigtest,"Study.ref"])

  # Standard deviations, as computed above
  sx <- filter(qPCR_time_course.summary,BCS==x1 & ExposureTime==timei &
Study==study.x)$deltaCt.RNU6.sd
  sy<-filter(qPCR_time_course.summary,BCS==y1 & ExposureTime==timei &
Study==study.y)$deltaCt.RNU6.sd

  # delTact mean
  mx<-filter(qPCR_time_course.summary,BCS==x1 & ExposureTime==timei &
Study==study.x)$deltaCt.RNU6
  my<-filter(qPCR_time_course.summary,BCS==y1 & ExposureTime==timei &
Study==study.y)$deltaCt.RNU6

  ## Compute t statistic/p value using mean, s.d. and sample size for two
groups
  tes<-t.test2(m1 = mx,
               m2= my,
               s1=sx,
               s2=sy,
               n1=3,
               n2=3)

```

```

qpcr_sigtest[sigtest,"pval"]<-tes["p-value"]
qpcr_sigtest[sigtest,"mean_diff"]<-tes["Difference of means"]

qpcr_sigtest[sigtest,"se"]<-sqrt(sy^2+sx^2)

## ddCt >0 implies x transcript is less abundant than y and ddCt<0 implies
x is more abundant than y. The relative magnitude on a plot of 2^(-x) !=
2(x) although the interpretability is the same (transcript is either 2^-x
less abundant than y or 2^x more abundant than y). Therefore, take absolute
ddCt before raising it to 2.
##Note ddct is actually 2^-ddct

if(-qpcr_sigtest[sigtest,"mean_diff"]<0){

  qpcr_sigtest[sigtest,"ddct"]<-(-2^(tes["Difference of means"]))
  qpcr_sigtest[sigtest,"ddct_se"]<-(-2^(tes["Difference of means"]+
  qpcr_sigtest[sigtest,"se"]))
  qpcr_sigtest[sigtest,"ddct_se_m"]<-(-2^(tes["Difference of means"]-
  qpcr_sigtest[sigtest,"se"]))
}else{

  qpcr_sigtest[sigtest,"ddct"]<-2^(-tes["Difference of means"])
  qpcr_sigtest[sigtest,"ddct_se"]<-2^((-tes["Difference of means"])+
  qpcr_sigtest[sigtest,"se"])
  qpcr_sigtest[sigtest,"ddct_se_m"]<-2^((-tes["Difference of means"])-
  qpcr_sigtest[sigtest,"se"])

}
}

## Annotate sig tests
qpcr_sigtest[qpcr_sigtest$pval<=pval_threshold &
!is.na(qpcr_sigtest$pval),"sig"]<-"*"

##convert exposure time to numeric for plot
qpcr_sigtest$ExposureTime.n<-
as.numeric(gsub(qpcr_sigtest$ExposureTime,pattern = "hr",replacement = ""))

qpcr_sigtest$ref<-as.character(qpcr_sigtest$reference)
qpcr_sigtest$comp<-as.character(qpcr_sigtest$comparison)

## Assign line types
qpcr_sigtest[qpcr_sigtest$ref=="G. vaginalis","lt"]<-1#"dashed"
qpcr_sigtest[qpcr_sigtest$ref=="Cell Culture Medium","lt"]<-3#"solid"
cairo_ps(paste0(R_script_output_directory,"miR_long_plot_presentation.eps"),w
idth = 11.5,height = 6)
## Plot qPCR Timecourse (put lactic acid comparison in sep. figure)

```

```

PLOT_QPCR_TIMECOURSE<-
  ggplot(filter(qpcr_sigtest,reference=="G. vaginalis" &
!grepl(qpcr_sigtest$comparison,pattern = "_10|_766|0_")), #&
qpcr_sigtest$reference=="G. vaginalis"),
  aes(x=ExposureTime.n,y=-mean_diff,col=comp))+
  theme_bw()+
  geom_hline(yintercept = 0,col="#9e9ac8",size=2.5)+
  theme(plot.margin=unit(c(0,0,0,0),units = "pt"),
text=element_text(size=16))+
  geom_line(size=1.5,show.legend = T,aes(linetype=ref))+
  geom_errorbar(aes(ymax=-mean_diff+se,ymin=-mean_diff-se),size=1,show.legend
= F)+
  geom_point(size=6,show.legend = F)+
  ylab(expression(paste(-Delta,Delta,"Ct")))+
  xlab("Exposure Time (hours)")+
  scale_color_manual(name=expression(paste("Lactobacillus
",Delta,"Ct")),values = color_scheme_BCS)+
  scale_x_continuous(breaks=c(0.5,1,4,11,13,22),minor_breaks = NULL)+
  scale_linetype_manual(name=expression(paste("Reference
",Delta,"Ct")),values=as.numeric(qpcr_sigtest$lt))+
  geom_text(aes(label=sig),col='#252525',size=8,nudge_y = -.05)

cairo_ps(paste0(thesis_figures_directory,FIGURE_QPCR_TIMECOURSE),width =
9,height = 6)
PLOT_QPCR_TIMECOURSE
dev.off()

## quartz_off_screen
##                2

qpcr_sigtest$comparison<-factor(qpcr_sigtest$comparison,levels = c("L.
crispatus", "L. jensenii","L. iners","G. vaginalis",
"PH_766","D_10","L_10","0_06_D","0_06_L"),ordered = T,labels = c("L.
crispatus", "L. jensenii","L. iners","G. vaginalis", "1% lactic acid, pH
7.66","0.1% D-lactic acid","0.1% L-lactic acid","0.06% D-lactic acid","0.06%
L-lactic acid"))

qpcr_sigtest$reference<-factor(qpcr_sigtest$reference,levels = c("L.
jensenii","L. iners","G. vaginalis", "Cell Culture Medium",
"L_10","0_06_D","0_06_L"),ordered = T,labels = c("L. jensenii","L.
iners","G. vaginalis", "Cell Culture Medium","0.1% L-lactic acid","0.06% D-
lactic acid","0.06% L-lactic acid"))

## Show D and L lactic acid qPCR results compared to other references
ddct_LacticAcid_plot<-
  ggplot(filter(qpcr_sigtest,ExposureTime=="4hr" &
!qpcr_sigtest$reference=="0.06% D-lactic acid"),aes(x=reference,y=-
mean_diff,col=comparison))+
  theme_bw()+

```

```

geom_hline(yintercept = 0,col="#9e9ac8",size=2.5)+
theme(plot.margin=unit(c(0,0,0,0),units = "pt"),
text=element_text(size=14))+
geom_errorbar(aes(ymax=-mean_diff+se,ymin=-mean_diff-
se),width=.2,show.legend = F,position = position_dodge(width = .2))+
geom_point(size=6,show.legend = T,position = position_dodge(width = .2))+
ylab(expression(paste(-Delta,Delta,"Ct")))+
xlab(expression(paste("Reference ",Delta,"Ct")))+
scale_color_manual(name=expression(paste("Exposure ",Delta,"Ct")),values =
c(color_scheme_BCS,"0.1% D-lactic acid"="#4dac26","0.1% L-lactic
acid"="#d01c8b","1% lactic acid, pH 7.66"="#386cb0","0.06% L-lactic
acid"="#f1b6da","0.06% D-lactic acid"="#b8e186"))+
geom_text(aes(label=sig),col='#252525',size=8,nudge_y = -.05)

cairo_ps(paste0(thesis_figures_directory,FIGURE_DL_LACTICACID_QPCR),width =
9,height = 6)
ddct_LacticAcid_plot
dev.off()

## quartz_off_screen
##                2

paste0("Additionally, the  $\Delta\Delta\text{Ct}$  of miR-193b expression after 4 hours of
exposure to 0.06% D-lactic acid relative to 0.06% L-lactic acid was found to
be non-significant ( p=",qpcr_sigtest[qpcr_sigtest$comparison=="L-lactic
acid" & qpcr_sigtest$reference=="D-lactic acid"],"pval"],").")

## [1] "Additionally, the  $\Delta\Delta\text{Ct}$  of miR-193b expression after 4 hours of
exposure to 0.06% D-lactic acid relative to 0.06% L-lactic acid was found to
be non-significant ( p=)."

## Write plot as a table
qpcr_sigtests<-
dplyr::select(qpcr_sigtest[with(qpcr_sigtest,order(ExposureTime.n,comparison,
reference)),],c(comparison,reference,ExposureTime,mean_diff,pval,sig))
qpcr_sigtests$Figure<-gsub(FIGURE_QPCR_TIMECOURSE,pattern=".eps",replacement
= "")

write.csv(qpcr_sigtests,file=paste0(thesis_tables_directory,TABLE_QPCR_TIMECO
URSE),row.names = F)

```

Quantify VK2 proliferation (function of miR-193b) exposed to BCS

- Expose VK2 cells to BCS
- Measure EdU detection and filled scratch area

```

#in_vitro_experiments<-read.csv(paste0("~/Dropbox
(IGS)/Jacques_Steve_Shared/Manuscript/miRNA/Working/R_script_input/", "In_vitr
o_Experiments.csv"))
#save(in_vitro_experiments,file=paste0(R_script_input_directory, "In_Vitro_Exp
eriments.Rdata"))

```

```

load(file=paste0(R_script_input_directory,"In_Vitro_Experiments.Rdata"))

## ////////////////
### Scratch Assay
## ////////////////

## Melt scratch assay data for easier handling
scratch.m<-filter(in_vitro_experiments,Experiment=="Scratch" & Observation ==
"Proliferation" & !grepl(x =
as.character(in_vitro_experiments$Treatment),pattern = "CONTROL_")) %>%
dplyr::select(-c(Experiment,Coverslip)) #melt(proliferation)

## Re-name the melted data
names(scratch.m)<-c("Observation","BCS","percent_cells","Field")

#Factor BCS so that order is enforced
scratch.m$BCS<-factor(scratch.m$BCS,ordered = T,levels = c("L. crispatus","L.
jensenii","L. iners","G. vaginalis","Cell Culture Medium"))

## Create a significance testing data frame to hold results
setup_sigtest.data<-setup_sigtest(pval_threshold = pval_threshold,raw_data =
scratch.m,test_function = "t.test",Experiment = "Scratch")

## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10

scratch_sigtest<-setup_sigtest.data$sigtest
scratch.summary<-setup_sigtest.data$summary_stats

## Create Barplots for scratch assay

plot_scratch.prolif<-

ggplot(scratch.summary)+geom_bar(aes(x=as.factor(BCS),y=mean,fill=BCS,col=BCS
),stat="identity",show.legend = F)+
  geom_errorbar(aes(x=BCS,ymin=mean-sd,ymax=mean+sd),width=.3,show.legend =
F)+
  mBio+
  ylab("Filled Scratch Area (%)")+

```

```

xlab("BCS")+
  theme(plot.margin=unit(c(0,0,0,0),units = "pt"),
text=element_text(size=16),axis.text.x=element_text(face="italic",angle =
45,vjust = 1,hjust = 1))+
  scale_fill_manual(values=color_scheme_BCS)+
  scale_y_continuous(limits=c(0,102))+
  scale_colour_manual(values=color_scheme_BCS)+

annotate("text",x=scratch_sigtest$midpoints,y=scratch_sigtest$y.sig,label=scr
atch_sigtest$sig,size=4)+

geom_segment(data=setup_sigtest.data$statbars,aes(x=x,xend=xend,y=y,yend=yend
),size=1)+

geom_segment(data=setup_sigtest.data$statbars2,aes(x=x,xend=xend,y=y,yend=yen
d),size=1)

##Write plot

cairo_ps(paste0(thesis_figures_directory,FIGURE_SCRATCH_QUANT),width =
8,height = 5.5)
plot(plot_scratch.prolif)
dev.off()

## quartz_off_screen
##          2

## ///////////////
### EdU Assay
## ///////////////

## Melt EdU assay data for easier handling, clean up input
EdU.m<-filter(in_vitro_experiments,Experiment=="Scratch" & Observation ==
"EdU") %>% dplyr::select(-c(Experiment)) #melt(scratch)
EdU.m<-rename(EdU.m,BCS=Treatment)

#Factor BCS so that order is enforced
EdU.m$BCS<-factor(EdU.m$BCS,ordered = T,levels = c("L. crispatus","L.
jensenii","L. iners","G. vaginalis","Cell Culture Medium"))

## Create a significance testing data frame to hold results
setup_sigtest.data<-setup_sigtest(pval_threshold = pval_threshold,raw_data =
EdU.m,test_function = "t.test",Experiment = "Scratch")

## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6

```

```

## [1] 7
## [1] 8
## [1] 9
## [1] 10

statbars<-setup_sigtest.data$statbars
statbars2<-setup_sigtest.data$statbars2
EdU_sigtest<-setup_sigtest.data$sigtest
EdU.summary<-setup_sigtest.data$summary_stats

## Create Barplots for EdU assay

EdUPlot<-

ggplot(EdU.summary)+geom_bar(aes(x=factor(BCS),y=mean,fill=BCS,col=BCS),stat=
"identity", position="dodge",show.legend = F)+
  geom_errorbar(aes(x=factor(BCS),ymin=mean-
sd,ymax=mean+sd),position="dodge",show.legend = F,width=.3)+
  mBio+
  ylab("Epithelial Cells Positive for EdU (%)")+
  xlab("BCS")+
  theme(plot.margin=unit(c(0,0,0,0),units = "pt"),
text=element_text(size=16),axis.text.x=element_text(face="italic",angle =
45,vjust = 1,hjust = 1))+
  scale_fill_manual(values=color_scheme_BCS)+
  scale_y_continuous(limits=c(0,102))+
  scale_colour_manual(values=color_scheme_BCS)+

annotate("text",x=EdU_sigtest$midpoints,y=EdU_sigtest$y.sig,label=EdU_sigtest
$y.sig,size=4)+
  geom_segment(data=statbars,aes(x=x,xend=xend,y=y,yend=yend),size=1)+
  geom_segment(data=statbars2,aes(x=x,xend=xend,y=y,yend=yend),size=1)

##Write plot

cairo_ps(paste0(thesis_figures_directory,FIGURE_EDU_QUANT),width = 8,height =
5.5)
plot(EdUPlot)
dev.off()

## quartz_off_screen
##                2

## Scratch assay data for TSB and NYC media exposed cells
control_meds<-filter(in_vitro_experiments,grepl(x =
as.character(in_vitro_experiments$Treatment),pattern = "CONTROL_"))

control_meds.nyc.mean<-
mean(control_meds[control_meds$Treatment=="CONTROL_NYCmed_20pct", "percent_cel
ls"])

```

```

control_meds.tsb.mean<-
mean(control_meds[control_meds$Treatment=="CONTROL_TSBmed_20pct", "percent_cells"])

control_meds.nyc.sd<-
sd(control_meds[control_meds$Treatment=="CONTROL_NYCmed_20pct", "percent_cells"])
control_meds.tsb.sd<-
sd(control_meds[control_meds$Treatment=="CONTROL_TSBmed_20pct", "percent_cells"])

control_meds.sig<-
t.test(x=filter(control_meds, Treatment=="CONTROL_NYCmed_20pct")[, "percent_cells"], y
=filter(control_meds, Treatment=="CONTROL_TSBmed_20pct")[, "percent_cells"], paired = F)

paste0("Percent cell proliferation between 20% NYC-III and TSB bacterial
culture media as evaluated by scratch assay was not significant (mean +/-
standard deviation NYC-III cell proliferation=", control_meds.nyc.mean, "% +/-
", control_meds.nyc.sd, "%, mean +/- standard deviation TSB cell
proliferation=", control_meds.tsb.mean, "% +/- ", control_meds.tsb.sd, "%,
p=", control_meds.sig$p.value, ")")

## [1] "Percent cell proliferation between 20% NYC-III and TSB bacterial
culture media as evaluated by scratch assay was not significant (mean +/-
standard deviation NYC-III cell proliferation=45.94945353125% +/-
9.11375283414706%, mean +/- standard deviation TSB cell
proliferation=,50.704530715% +/- 14.5706994714208%, p=0.449364642797794)"

## Combine proliferation sig tests and write to file
proliferation_sigtests<-
rbind(data.frame(Figure=gsub(Figure_SCRATCH_QUANT, replacement = "", pattern =
".eps"), scratch_sigtest), data.frame(Figure=gsub(Figure_EDU_QUANT, replacement =
"", pattern = ".eps"), Edu_sigtest))
proliferation_sigtests<-proliferation_sigtests[with(proliferation_sigtests,
order(Figure, x, xend)),]
proliferation_sigtests<-
dplyr::select(proliferation_sigtests, Figure, xref, reference, pval, mean_diff, sig
)

proliferation_sigtests<-
rbind(proliferation_sigtests, data.frame(Figure="FIGURE_S5", xref="TSB", referen
ce="NYC", pval=control_meds.sig$p.value, mean_diff=unname(control_meds.sig$esti
mate[2])-
control_meds.sig$estimate[1]), sig=ifelse(control_meds.sig$p.value<pval_thresh
old, "*", "N.S.)))

```

```

PLOT_nyc_v_tsb<-ggplot(data.frame(media=c("NYC (Lactobacillus spp.)", "TSB (G.
vaginalis)"), scratch.mean=c(control_meds.nyc.mean, control_meds.tsb.mean), scratch.sd=c(control_meds.nyc.sd, control_meds.tsb.sd)))+
  geom_bar(aes(x=media, y=scratch.mean, fill=media), stat="identity", show.legend=F)+
  geom_errorbar(aes(x=media, ymax=scratch.mean+scratch.sd, ymin=scratch.mean-
scratch.sd), width=.2)+
  mBio+ylab("Percent Scratch Area Filled")+
  xlab("VK2 Cell Culture Medium + 20% Bacterial Culture Medium")+
  scale_y_continuous(limits=c(0,100))+
  annotate("text", x=1.5, y=90, label="N.S.", size=4)+

geom_segment(data=data.frame(x=1, xend=2, y=87, yend=87), aes(x=x, xend=xend, y=y, y
end=yend), size=1)+

geom_segment(data=data.frame(x=c(1,2), xend=c(1,2), y=c(85,85), yend=c(87,87)), a
es(x=x, xend=xend, y=y, yend=yend), size=1)

cairo_ps(file=paste0(thesis_figures_directory, FIGURE_nyc_v_tsb), width =
8, height = 5.5)
PLOT_nyc_v_tsb
dev.off()

## quartz_off_screen
##          2

write.csv(proliferation_sigtests, file=paste0(thesis_tables_directory, TABLE_ED
U_SCRATCH_QUANT), row.names = F)

## PErcent proliferation in 1% DL lactic acid, ph buffered 7.66 expoed cells
DL_PH766_percent_proliferation<-data.frame(DL_1pct_766=c(
7.706708323,
4.631917974,
8.766697901,
7.771365876)
)
mean(DL_PH766_percent_proliferation$DL_1pct_766)

## [1] 7.219173

sd(DL_PH766_percent_proliferation$DL_1pct_766)

## [1] 1.791771

t.test(x = scratch.m[scratch.m$BCS=="Cell Culture Medium", "percent_cells"], y
= DL_PH766_percent_proliferation$DL_1pct_766)

##
## Welch Two Sample t-test
##
## data:  scratch.m[scratch.m$BCS == "Cell Culture Medium", "percent_cells"]

```

```

and DL_PH766_percent_proliferation$DL_1pct_766
## t = 6.0479, df = 2.0264, p-value = 0.02546
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 19.92453 114.08363
## sample estimates:
## mean of x mean of y
## 74.223251 7.219173

t.test(x = scratch.m[scratch.m$BCS=="G. vaginalis", "percent_cells"], y =
DL_PH766_percent_proliferation$DL_1pct_766)

##
## Welch Two Sample t-test
##
## data: scratch.m[scratch.m$BCS == "G. vaginalis", "percent_cells"] and
DL_PH766_percent_proliferation$DL_1pct_766
## t = 6.0415, df = 2.2459, p-value = 0.01982
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 8.070304 37.070274
## sample estimates:
## mean of x mean of y
## 29.789462 7.219173

```

Inhibit A2EN cell proliferation & Observe CT Infectivity

- Expose A2EN cells to proliferation inhibitors
- Infect w/ CT
- Measure EdU detection and infection

```

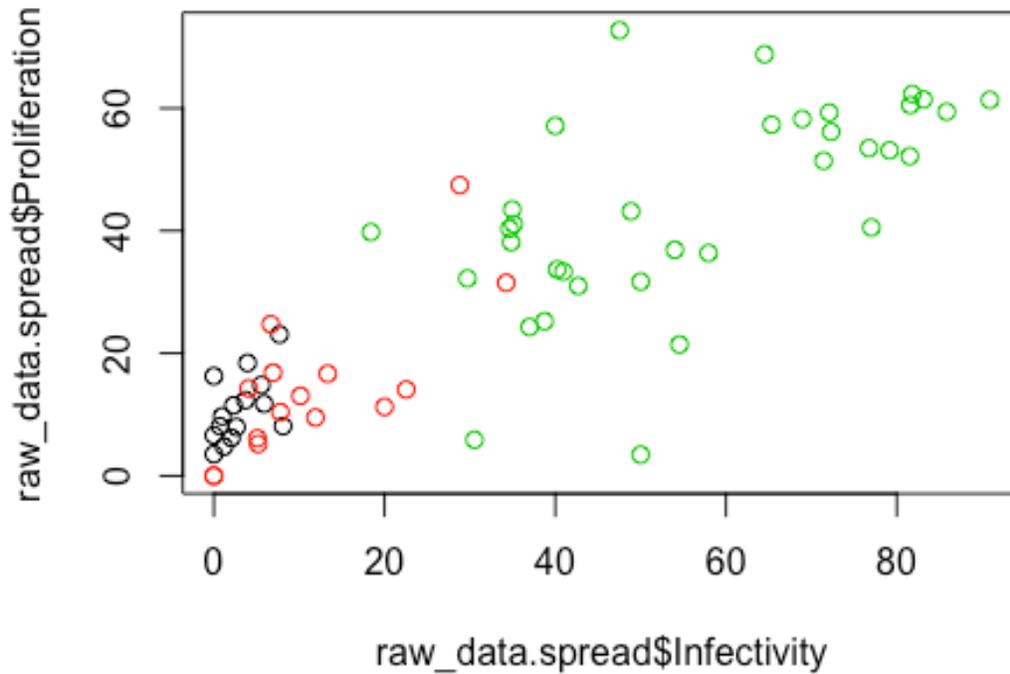
load(file=paste0(R_script_input_directory, "In_Vitro_Experiments.Rdata"))

## Melt Ct infection assay data for easier handling
pi.m<-filter(in_vitro_experiments, Experiment=="Infection") %>%
dplyr::select(-Experiment)

## Factor Treatment to enforce order and make labels
pi.m$Treatment<-factor(pi.m$Treatment, levels = c("Cdk4 400nM", "Fascaplysin
350nM", "Cell Culture Medium"), ordered = T, labels = c("CAS 546102-60-
7", "Fascaplysin", "Cell Culture Medium"))

## Create a significance testing data frame to hold results
setup_sigtest.data<-setup_sigtest(pval_threshold = pval_threshold, raw_data =
pi.m, test_function = "t.test2", Experiment = "Infection")

```



```
## [1] 1
## [1] "Proliferation"
## Difference of means          Std Error          t
##      -6.6236985          3.8807666          -1.7068016
##      p-value
##      0.1488428
## [1] 2
## [1] "Proliferation"
## Difference of means          Std Error          t
##      -3.515431e+01      4.999389e+00      -7.031721e+00
##      p-value
##      6.107794e-05
## [1] 3
## [1] "Proliferation"
## Difference of means          Std Error          t
##      -2.853061e+01      5.386636e+00      -5.296555e+00
##      p-value
##      3.539033e-04
## [1] 4
## [1] "Infectivity"
## Difference of means          Std Error          t
##      -8.79632006      2.06013167      -4.26978536
##      p-value
```

```

##           0.01525339
## [1] 5
## [1] "Infectivity"
## Difference of means           Std Error           t
##      -5.338663e+01           7.813379e+00           -6.832720e+00
##           p-value
##           3.346066e-04
## [1] 6
## [1] "Infectivity"
## Difference of means           Std Error           t
##      -44.590306829           7.736338543           -5.763748132
##           p-value
##           0.001000266

pi.summary<-setup_sigtest.data$summary_stats
ct_sigtest<-setup_sigtest.data$sigtest
statbars<-setup_sigtest.data$statbars
statbars2<-setup_sigtest.data$statbars2

## Split data into infection and proliferation (EdU) data
ct_sigtest_inf<-filter(ct_sigtest,Observation=="Infectivity")
ct_sigtest_pro<-filter(ct_sigtest,Observation=="Proliferation")

## Plot Infectivity
Fig3_inf<-
ggplot(filter(pi.summary,Observation=="Infectivity"),aes(x=Treatment,y=grand_mean))+geom_bar(aes(fill=Treatment),stat="identity",show.legend =
F)+geom_errorbar(aes(ymin=grand_mean-
grand_sd,ymax=grand_mean+grand_sd,width=.3))+
  scale_fill_manual(values=c("Cell Culture
Medium"='blue',"Fascaplysin"="#fdb863","CAS 546102-60-7"="#b2df8a"))+
  mBio+
  theme(plot.margin=unit(c(0,0,0,0),units = "pt"),
text=element_text(size=16),axis.text.x=element_text(angle = 45,vjust =
1,hjust = 1,size=20))+scale_y_continuous(limits=c(0,102))+

annotate("text",x=ct_sigtest_inf$midpoints,y=ct_sigtest_inf$y.sig,label=ct_si
gtest_inf$y.sig,size=2.5)+
  geom_segment(data=statbars,aes(x=x,xend=xend,y=y,yend=yend))+

geom_segment(data=statbars2,aes(x=x,xend=xend,y=y,yend=yend),col="Black")+ylab
("Epithelial Cells with C. trachomatis Inclusion (%)")

## Write plot

cairo_ps(paste0(thesis_figures_directory,FIGURE_CT_INFECT_QUANT),width=6,heig
ht = 5.5)#,width = 8,height = 5.5)
plot(Fig3_inf)
dev.off()

```

```

## quartz_off_screen
##           2

##Plot Ct Assay Edu

Fig3_prolif<-
ggplot(filter(pi.summary,Observation=="Proliferation"),aes(x=Treatment,y=grand_mean))+geom_bar(stat="identity",show.legend =
F,aes(fill=Treatment))+geom_errorbar(aes(ymin=grand_mean-
grand_sd,ymax=grand_mean+grand_sd,width=.3))+mBio+scale_fill_manual(values=c(
"Cell Culture Medium"='blue',SIS3="#c2a5cf","Fascaplysin"="#fdb863","CAS
546102-60-7"="#b2df8a"))+ylab("Epithelial Cells Positive for Edu (%)")+
  theme(plot.margin=unit(c(0,0,0,0),units = "pt"),
text=element_text(size=20),axis.text.x=element_text(angle = 45,vjust =
1,hjust = 1))+scale_y_continuous(limits=c(0,102))+

annotate("text",x=ct_sigtest_pro$midpoints,y=ct_sigtest_pro$y.sig,label=ct_si
gtest_pro$y.sig,size=2.5)+
  geom_segment(data=statbars,aes(x=x,xend=xend,y=y,yend=yend))+
  geom_segment(data=statbars2,aes(x=x,xend=xend,y=y,yend=yend),col="Black")

##Write plot
cairo_ps(paste0(thesis_figures_directory,FIGURE_CT_EDU_QUANT),width=6,height
= 5.5)#width = 8,height = 5.5)
plot(Fig3_prolif)
dev.off()

## quartz_off_screen
##           2

## Print text describing results

## Create variables to hold values for printing
ct_sigtest<-unique(ct_sigtest)
pro_C<-ct_sigtest[ct_sigtest$Observation=="Proliferation" &
ct_sigtest$reference=="Cell Culture Medium" & ct_sigtest$xref=="CAS 546102-
60-7" ,c("mean_diff","pval")]
pro_F<-ct_sigtest[ct_sigtest$Observation=="Proliferation" &
ct_sigtest$reference=="Cell Culture Medium" & ct_sigtest$xref=="Fascaplysin"
,c("mean_diff","pval")]

ct_C<-ct_sigtest[ct_sigtest$Observation=="Infectivity" &
ct_sigtest$reference=="Cell Culture Medium" & ct_sigtest$xref=="CAS 546102-
60-7" ,c("mean_diff","pval")]
ct_F<-ct_sigtest[ct_sigtest$Observation=="Infectivity" &
ct_sigtest$reference=="Cell Culture Medium" & ct_sigtest$xref=="Fascaplysin"
,c("mean_diff","pval")]

paste0("Epithelial cell proliferation was decreased by ",-pro_C$mean_diff,"%

```

```

(p=",pro_C$pval,") in CAS 546102-60-7 and ",-pro_F$mean_diff,"%
(p=",pro_F$pval,") in Fascaplysin treated cells relative to Cell Culture
Medium treated cells, respectively")

## [1] "Epithelial cell proliferation was decreased by 35.1543095335083%
(p=6.10779413826187e-05) in CAS 546102-60-7 and 28.530611023625%
(p=0.000353903348334145) in Fascaplysin treated cells relative to Cell
Culture Medium treated cells, respectively"

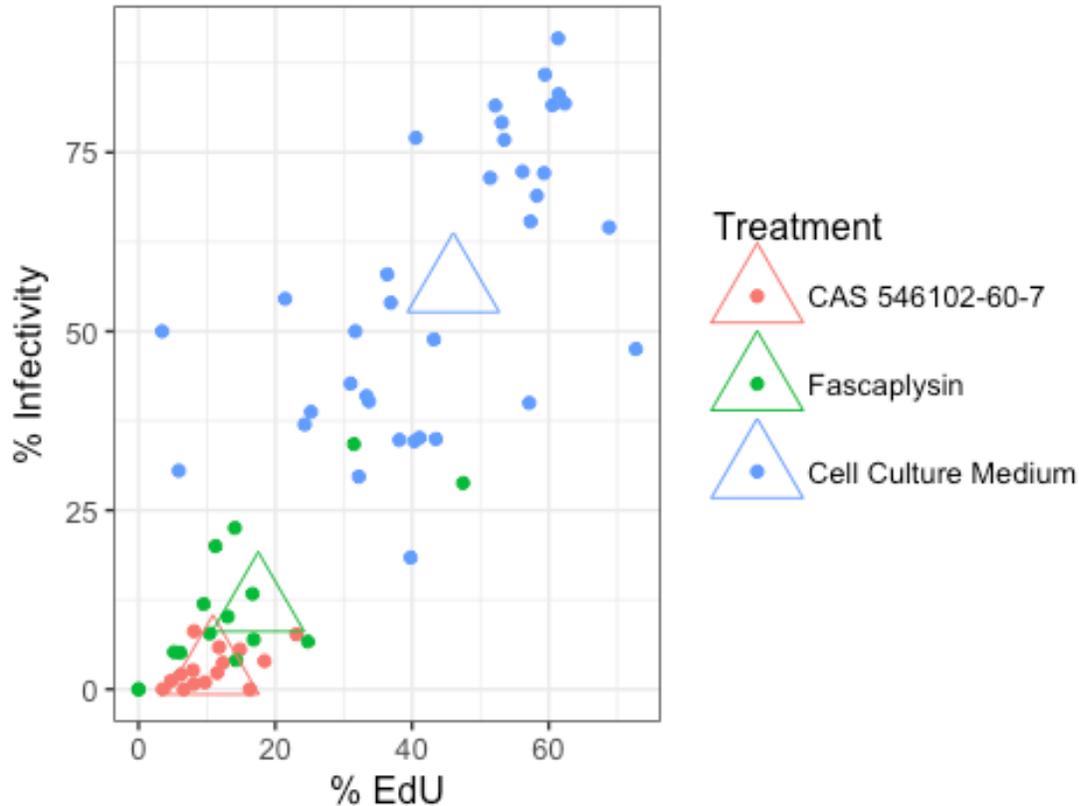
paste0("C. trachomatis infection was decreased by ",-ct_C$mean_diff,"%
(p=",ct_C$pval,") in CAS 546102-60-7 and ",-ct_F$mean_diff,"%
(p=",ct_F$pval,") in Fascaplysin treated cells relative to Cell Culture
Medium treated cells, respectively")

## [1] "C. trachomatis infection was decreased by 53.3866268869905%
(p=0.000334606564696431) in CAS 546102-60-7 and 44.5903068287238%
(p=0.00100026580119664) in Fascaplysin treated cells relative to Cell Culture
Medium treated cells, respectively"

## Calcualte correlation between mean infectivity and mean proliferation
corr_mx<-spread(pi.m,Observation, value=percent_cells)
ggplot(corr_mx,aes(y=Infectivity,x=Proliferation,col=Treatment))+geom_point()
+ggtitle("Ct Infectivity vs EdU")+mBio+ylab("% Infectivity")+xlab("%
EdU")+geom_point(data=spread(dplyr::select(pi.summary,-
c(grand_sd,n)),Observation,grand_mean),aes(y=Infectivity,x=Proliferation,col=
Treatment),size=10,pch=2)

```

Ct Infectivity vs EdU



```
ct_pro_lm<-lm(formula = Infectivity~Proliferation,data =
data.frame(corr_mx[!rowSums(is.na(corr_mx))>0,]))

ct_pro_summary_lm<-
lm(Infectivity~Proliferation,data=spread(dplyr::select(pi.summary, -
c(grand_sd,n)),Observation,grand_mean))

summary(ct_pro_lm)

##
## Call:
## lm(formula = Infectivity ~ Proliferation, data =
data.frame(corr_mx[!rowSums(is.na(corr_mx)) >
##   0, ]))
##
## Residuals:
##   Min     1Q   Median     3Q      Max
## -37.707  -9.275  -1.017   9.703  47.669
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.79531    3.29799  -0.544   0.588
## Proliferation  1.19670    0.09104  13.145 <2e-16 ***
```

```

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.21 on 63 degrees of freedom
## Multiple R-squared:  0.7328, Adjusted R-squared:  0.7286
## F-statistic: 172.8 on 1 and 63 DF,  p-value: < 2.2e-16

summary(ct_pro_summary_lm)

##
## Call:
## lm(formula = Infectivity ~ Proliferation, data =
spread(dplyr::select(pi.summary,
##      -c(grand_sd, n)), Observation, grand_mean))
##
## Residuals:
##      1      2      3
## 0.6049 -0.7453  0.1404
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.27998   1.06922  -13.36  0.0476 *
## Proliferation  1.53185   0.03672   41.72  0.0153 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9701 on 1 degrees of freedom
## Multiple R-squared:  0.9994, Adjusted R-squared:  0.9989
## F-statistic: 1740 on 1 and 1 DF,  p-value: 0.01526

## Combine proliferation sig tests and write to file
ct_sigtests<-
data.frame(Figure=paste0("FIGURE_4",ct_sigtest$Observation),ct_sigtest)
ct_sigtests<-
dplyr::select(ct_sigtests[with(ct_sigtests,order(Observation,x,xend)),],c(Fig
ure,xref,reference,pval,mean_diff,sig))

write.csv(ct_sigtests,file=paste0(thesis_tables_directory,TABLE_CT_QUANT),row
.names=F,quote=F)

```

Subject Longitudinal Plots (Figure 1)

Create longitudinal plots for subjects used in study

```

### Load previously prepared 16S metataxonomic data. metatdata
load(file=paste0(R_script_input_directory,"subject_plot_data.Rdata"))
OTU_METADATA<-subject_plot_data$OTU_METADATA ## metadata
rRNA_16S<-subject_plot_data$relativeAbundance ## taxa assignments/relative
abundances
sampleInfoColNames<-subject_plot_data$sampleInfoColNames ## holds which

```

```

column names are associated with metadata

miRNA_extractions<-SRL_meta_table[is.na(SRL_meta_table$QC_removal_stage),] ##
Samples associated with miRNA libraries used in final study- i.e., don't
include samples removed due to poor QC.
subject_plot_list<- unique(miRNA_extractions$SID) ## Subject IDs to plot

global_species_list<-NULL ## Initialize a container to store all species
plotted in fig 1 as a legend

for(s in subject_plot_list){ ## iterate through the subject's
  print(s)
  #s<- "UAB008"
  ## ////////////////////////////////////// ##
  ## /// Subeset Data by Subject /// ##
  ## ////////////////////////////////////// ##

  relabundance<-filter(rRNA_16S,SID==s) ## taxa relative abundance for
subject
  otu_count<-OTU_METADATA[OTU_METADATA$SID==s,] ## Metadata for subject

  ## Drop 16S samples that have less than the threshold for high confidence
taxa assignments
  lowCountThreshold<-1000
  low_count_samples<-
otu_count[otu_count$X16S_total_counts<lowCountThreshold,"SERIAL"]

  ##Nugent score data for subject
  nugent<-filter(OTU_METADATA,SID==s) %>% dplyr::select(SERIAL,NUGENT_SCORE)
  nugent$color<-"grey"
  ## Colors/handling for missing Nugent scores
  nugent[is.na(nugent$NUGENT_SCORE),"color"]<-"red"
  nugent[!is.na(nugent$NUGENT_SCORE),"color"]<-"black"
  nugent[is.na(nugent$NUGENT_SCORE),"NUGENT_SCORE"]<-(-1)

  ## pH data for subject
  ph<-filter(OTU_METADATA,SID==s) %>% dplyr::select(SERIAL,PH)
  ## Colors/handling for missing pH values
  ph$color<-"grey"
  ph[is.na(ph$PH),"color"]<-"red"
  ph[!is.na(ph$PH),"color"]<-"black"
  ph[is.na(ph$PH),"PH"]<-2

  ##METadata for subject
  dailyDiaryMetadata<-filter(OTU_METADATA,SID==s)

  ## ////////////////////////////////////// ##
  ## Determine global plot time scale ##
  ## ////////////////////////////////////// ##

```

```

SERIAL_global<-
sort(unique(dailyDiaryMetadata$SERIAL,relabundance$SERIAL,miRNA_extractions$SERIAL)) ## All time points
day<-SERIAL_global%%7 ## Day relative to all SERIALized time points
day[SERIAL_global%%7==0]<-7 ## the mod calculation causes all day 7 to be 0. Repalce w/ 7.
week<-((SERIAL_global-day)/7)+1 ## Week back calculated from SERIALized time points.
time_points<-data.frame(SERIAL=SERIAL_global, day=day, week=week) ## container
time_points$plot_label=""
time_points[time_points$day==7,"plot_label"]<-
time_points[time_points$day==7,"week"] ## plot labels every week
global_min_time<-
max(c(min(miRNA_extractions[miRNA_extractions$SID==s,"SERIAL"]-5),min(SERIAL_global,na.rm = T) ))

## min time point - the max of the min of either miR extractions or all data
global_max_time<-
min(c(max(miRNA_extractions[miRNA_extractions$SID==s,"SERIAL"]+5),max(SERIAL_global,na.rm = T))) ## max time point across all data
removed_samples
## cleaner way to define as variable for X axis:
time_breaks<-time_points$SERIAL
time_label<-time_points$plot_label
time_limits<-c(global_min_time-1,global_max_time+1)

## Determine the min & max of time points
rect_min<-time_breaks[!time_breaks%in%
miRNA_extractions[miRNA_extractions$SID==s,"SERIAL"]]-.5
rect_max<-time_breaks[!time_breaks%in%
miRNA_extractions[miRNA_extractions$SID==s,"SERIAL"]]+.5

## //////////////////////////////////////////////////////////////////// ##
## / Determine most abundant species ##
## //////////////////////////////////////////////////////////////////// ##

# ////////////////////////////////////////////////////////////////////
# Most abundant species defined per subject & based on cutoff. ##
# ALL other taxa binned into "other" #
# ////////////////////////////////////////////////////////////////////

### Grab just relative abundances, no sample info columns:
relabundance[is.na(relabundance)]<-0
relabundance<-relabundance[relabundance$SERIAL>=time_limits[1] &
relabundance$SERIAL<=time_limits[2],]
relabundance_for_max_calc<-relabundance[,!(names(relabundance) %in%

```

```

sampleInfoColNames)]

### Calucalte max for each taxa
max_relabundances<-apply(relabundance_for_max_calc,2,max)
### number of species whose rel abundance is above a certain threshold
numHighAbundSpecies<-sum(max_relabundances>raThreshold,na.rm = T)

##### Plot either the top X most abundant species,
##### or the most abundant species above max plottable species, whichever
is lesser.
##### This helps prevent "taxa overload" on the plot.

most_abundant_species<-"" ## Will hold names of most abundant species.
if(numHighAbundSpecies>nSpecies){
  most_abundant_species<-names(sort(max_relabundances,decreasing =
T))[1:nSpecies]
}else{
  most_abundant_species<- names(max_relabundances[!is.na(max_relabundances)
& max_relabundances>raThreshold])
}

##### Pull the most abundant species (defined above), bin the remainder
into "other".
##### Also recombine the 'sample info' onto most abundant/other species
table.

most_abundant_relabundance<-relabundance[most_abundant_species]
other_relabundance<-rowSums(relabundance[!names(relabundance) %in%
most_abundant_species |
names(relabundance) %in%
sampleInfoColNames)],na.rm=T)
otu_count_relabundance<-cbind(relabundance[names(relabundance) %in%
c("SERIAL",sampleInfoColNames)],
most_abundant_relabundance,Other=other_relabundance)

##Update global species list with any new species
global_species_list<-
unique(c(most_abundant_species,global_species_list,"Other"))

reshape_names<-c("Pre_QC_ID","SID","UID","SERIAL")
otu_count_reshape<-melt(data =
otu_count_relabundance,id.vars=reshape_names)
names(otu_count_reshape)<-c(reshape_names,"species","count")

### ///////////////////////////////////////// ###
### ////////////// OTU Plot ////////////// ###
### ///////////////////////////////////////// ###

```

```

otuPlot<- ggplot(otu_count_reshape)+
  geom_area(aes(x=SERIAL,y=count,fill=species),
            stat="identity",show.legend=F,position="fill")+#,width=1)+
  mBio+
  theme(axis.ticks = element_blank(),
        axis.title.x=element_blank(),
        axis.text.x = element_blank(),
        axis.text.y = element_text(size = rel(1.5)),
        legend.key.size=unit(8, "points"),
        legend.title = element_blank(),
        legend.text = element_text(size = rel(.5),face="italic"),
        axis.title = element_text(size = rel(sizes)),
        plot.margin=unit(c(2.5,40,2.5,10),units="points"),#margins
        panel.grid.major.y=element_line(colour = "grey73"),
        panel.grid.minor.x = element_blank()+
  ylab("Phylotype Relative\nAbundance (%)")+
  geom_vline(xintercept =
otu_count_reshape$SERIAL,size=rel(.2),col="grey")+
  scale_fill_manual(values=subject_long_taxa_colors)+
  ggtitle(paste0(s))+
  annotate("rect", xmin=rect_min, xmax=rect_max, ymin=0, ymax=1,
alpha=alpha_rect, fill=rect_fill)+

scale_x_continuous(breaks=time_breaks,label=time_label,limits=time_limits)

## Determine any dropped sample (post QC) time points and place an * above
dropped_samples.serial<-otu_count[otu_count$Pre_QC_ID %in%
removed_samples$Pre_QC_ID,"SERIAL"]
if(length(dropped_samples.serial)!=0){otuPlot<-otuPlot+annotate("text", x
=dropped_samples.serial , y = 1.01, label = "*",size=8)}

### //////////////////////////////////////// ###
### //////////          Nugent Plot          ////////// ###
### //////////////////////////////////////// ###
nugentPlot<-ggplot(data=nugent)+
  geom_bar(aes(x=as.numeric(SERIAL),y=NUGENT_SCORE,width=.9,fill=color),
            stat="identity",position = position_dodge(width=0.5))+
  mBio+
  geom_hline(yintercept = c(3,7),size=rel(1),col="pink")+ ## Defines Nugent
score 3 & 7 (BV
  theme(axis.ticks = element_blank(),
        axis.title.x=element_blank(),
        axis.text.x = element_blank(),
        axis.text.y = element_text(size = rel(1.5*sizes)),
        legend.position="none",
        plot.margin=margins,
        panel.grid.major.y=element_line(colour = "grey73"),

```

```

        panel.grid.minor.x = element_blank()+
        ylab("Nugent\nScore")+
        scale_fill_manual(values=c('black','red'))+

scale_x_continuous(breaks=time_breaks,label=time_label,limits=time_limits)+
  annotate("rect", xmin=rect_min, xmax=rect_max, ymin=0, ymax=10,
alpha=alpha_rect, fill=rect_fill)+
  scale_y_continuous(breaks=c(0,3,7,10),limits=c(-1.25,10.5))

### ///////////////////////////////////////// ###
### ///////////////////////////////////////// Metadata Plot ///////////////////////////////////////// ###
### ///////////////////////////////////////// ###
metaPlot<-ggplot(dailyDiaryMetadata, aes(x=as.numeric(SERIAL)))+
  geom_point(aes(y=1*(as.numeric(VAG_DIS)==1),size=2),
            pch=16,col='blue',position = position_dodge(width=0.5))+
  geom_point(aes(y=2*(as.numeric(VAG_ODOR)==1),size=2),
            pch=16,col='blue',position = position_dodge(width=0.5))+

geom_point(aes(y=3*(as.numeric(MENSTRUATION)>0),size=as.numeric(MENSTRUATION)
),
            col='red',pch=16, position = position_dodge(width=0.5))+
  scale_size(range = c(2+2,4+2))+ ## for menstru point sizes. Bounded by 3
points on a scale.
mBio+
  theme(legend.position="none",
        plot.title = element_text(size = rel(sizes)),
        axis.text = element_text(size = rel(2*sizes)),
        axis.title = element_text(size = rel(sizes)),
        axis.title.y=element_text(vjust=.2),
        axis.title.x=element_text(vjust=-.2),
        axis.text.y=element_text(size=rel(0.75)), ## change back to 0.75
        plot.margin=unit(c(-2.5,40,5,5),units="points"),
        panel.grid.major.y=element_line(colour = "grey73"),
        panel.grid.minor.x = element_blank()+
        xlab("Week")+
        ylab(""))+
  annotate("rect", xmin=rect_min, xmax=rect_max, ymin=1, ymax=3,
alpha=alpha_rect, fill=rect_fill)+

scale_x_continuous(breaks=time_breaks,label=time_label,limits=time_limits)+
  scale_y_continuous(breaks=1:3,labels = c("Discharge",
            "Odor",
"Menstruation"),limits=c(0.5,3.5))

##### Determine if subject has low qualiyt daily dairy flag and annotate
plot:

```

```

if(!sum(is.na(dailyDiaryMetadata$Diary_QUALITY_FLAG))){
  min<-min(timeTable$SERIAL,na.rm = T)
  max<-max(timeTable$SERIAL,na.rm = T)
  middle.x<-(max-min)/2
  middle.y<<-(16-1)/2
  metaPlot<-metaPlot + annotate("text", x = middle.x,
                                y = middle.y, label = "?",
                                size=rel(40),
                                col="grey")
}

### ///////////////////////////////////////// ###
### ///////////////////////////////////////// pH Plot ///////////////////////////////////////// ###
### ///////////////////////////////////////// ###
ph_normalization_factor<-3.5
## Notice the pH value scale is "normalized" by subtracting
"ph_normalization_factor" from the actual pH value, then re-labeling the y
axis. This is very dangerous, but ggplot will not permit bar plots that start
from > 0 .
ph[ph$PH==2,"PH"]<-ph_normalization_factor-0.5

phPlot<-ggplot(ph,aes(x=as.numeric(SERIAL),
                      y=as.numeric(PH)-
ph_normalization_factor,width=.9,fill=color))+
  geom_bar(stat="identity",position = position_dodge(width=0.5))+
  scale_fill_manual(values=c('black','red'))+
  mBio+
  geom_hline(yintercept = c(4.5-ph_normalization_factor),
             size=rel(1),col="pink")+ ## Vaginal pH>4.5 one criteria for
BV.
  theme(axis.ticks = element_blank(),
        axis.title.x=element_blank(),
        axis.text.x = element_blank(),
        axis.text = element_text(size = rel(1.5*sizes)),
        legend.position="none",
        plot.margin=margins,
        panel.grid.major.y=element_line(colour = "grey73"),
        panel.grid.minor.x = element_blank())+
  ylab("pH")+
  annotate("rect", xmin=rect_min, xmax=rect_max, ymin=min(0,min(ph$PH,na.rm
= T)-ph_normalization_factor-0.25), ymax= max(ph$PH,na.rm = T)+0.25-
ph_normalization_factor, alpha=alpha_rect, fill=rect_fill)+

scale_x_continuous(breaks=time_breaks,label=time_label,limits=time_limits)+
  scale_y_continuous(breaks=c(4,4.5,5,5.5)-ph_normalization_factor,
limits=c(-0.75,5.8+0.25-
ph_normalization_factor),labels=c("4","4.5","5","5.5"))

```

```

### ///////////////////////////////////////// ###
### //Tie Plots Together // ###
### ///////////////////////////////////////// ###

## ///////////////////////////////////////// ##
## //Define plots as Grobs// ##
## ///////////////////////////////////////// ##

grob.otuPlot <- ggplotGrob(otuPlot)
grob.nugentPlot <- ggplotGrob(nugentPlot)
grob.phPlot<-ggplotGrob(phPlot)
grob.metaPlot <- ggplotGrob(metaPlot)

## ///////////////////////////////////////// ##
## /// Find max width /// ##
## ///////////////////////////////////////// ##
maxWidth = grid::unit.pmax(grob.otuPlot$widths[1:6],
                           grob.nugentPlot$widths[1:5],
                           grob.phPlot$widths[1:5],
                           grob.metaPlot$widths[1:5])

## ///////////////////////////////////////// ##
## /Redefine common max width ##
## ///////////////////////////////////////// ##
grob.nugentPlot$widths[1:6] <- as.list(maxWidth)
grob.otuPlot$widths[1:6] <- as.list(maxWidth)
grob.metaPlot$widths[1:6] <- as.list(maxWidth)
grob.phPlot$widths[1:6]<-as.list(maxWidth)

### ///////////////////////////////////////// ###
### //Write/Draw Plot // ###
### ///////////////////////////////////////// ###

cairo_ps(paste0(thesis_figures_directory,FIGURE_SUBJECT_PLOTS,s,".eps"),width
= 11,height = 8.5)

grid.arrange(grob.otuPlot,
              grob.nugentPlot,
              grob.phPlot,
              grob.metaPlot,
              ncol=1,nrow=4,
              heights=c(2.5,1,1,1))

dev.off()
}

## [1] "EM12"

## [1] "UAB003"

```

```

## [1] "UAB005"
## [1] "UAB006"
## [1] "UAB007"
## [1] "UAB008"
## [1] "UAB015"
## [1] "UAB021"
## [1] "UAB022"
## [1] "UAB055"
## [1] "UAB093"
## [1] "UAB102"
## [1] "UAB115"
## [1] "UAB116"
## [1] "UAB117"
## [1] "UAB121"

## /////
## The following plots the figure legend containing colors for all taxa
plotted in for loos

dummy_globalSpeciesList<-
data.frame(SERIAL=1,species=global_species_list,count=1)
dummy_globalSpeciesList_plot<-ggplot(dummy_globalSpeciesList)+

geom_area(aes(x=SERIAL,y=count,fill=species,order=order(as.numeric(as.factor(
dummy_globalSpeciesList$species))),decreasing = F)),
          stat="identity",show.legend=T,position="fill")+#,width=1)+
mBio+
theme(axis.ticks = element_blank(),
      axis.title.x=element_blank(),
      axis.text.x = element_blank(),
      legend.key.size=unit(8, "points"),
      legend.title = element_blank(),
      legend.text = element_text(size = rel(.5),face="italic"),
      axis.title = element_text(size = rel(sizes)),
      plot.margin=unit(c(2.5,40,2.5,10),units="points"),#margins
      panel.grid.major.y=element_line(colour = "grey73"))+
scale_fill_manual(values=subject_long_taxa_colors)+
ggtitle("Global Species List Color Codes")

##Write legend to file

```

```

cairo_ps(paste0(thesis_figures_directory,FIGURE_SUBJECT_PLOTS,"GlobalSpeciesL
ist.eps"),width = 11,height = 8.5)
dummy_globalSpeciesList_plot
dev.off()

## quartz_off_screen
##                2

```

Ribo-reduced RNA-seq Analysis

```

## RIN Quality Distribution
TRL_RNA_Sample_QuantQual<-
read.csv(paste0(R_script_input_directory,"TRL_RNA_Sample_Quality.csv"),header
= F)
postscript(paste0(thesis_figures_directory,FIGURE_TRL_RIN_HIST))
p<-
ggplot(TRL_RNA_Sample_QuantQual,aes(x=V1))+geom_histogram()+xlab("RINe")+ylab
("Number of Samples")+geom_vline(xintercept =
median(TRL_RNA_Sample_QuantQual$V1),col='red')+mBio
plot(p)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

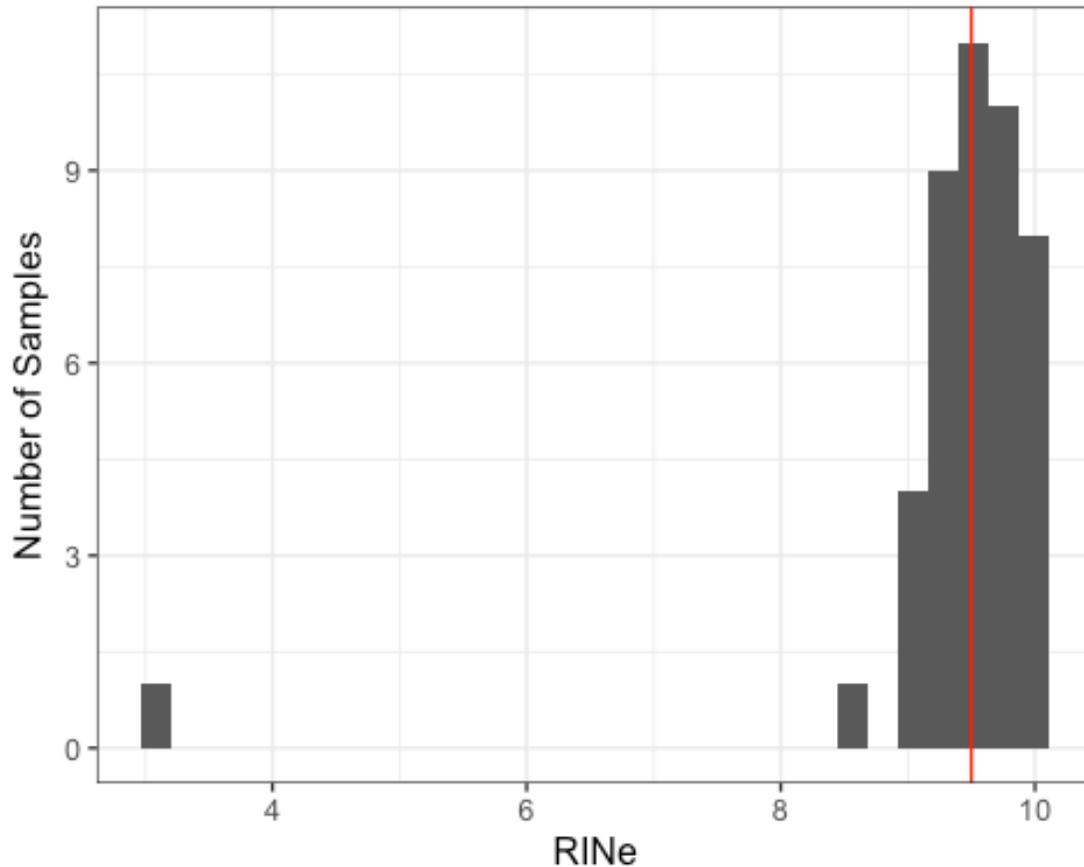
dev.off()

## quartz_off_screen
##                2

plot(p)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

```



```
summary(TRL_RNA_Sample_QuantQual$V1)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      3.100  9.200  9.500  9.366  9.800  10.000
```

```
summary(10*TRL_RNA_Sample_QuantQual[TRL_RNA_Sample_QuantQual$V1!=3.1,"V2"])
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      226.0  438.5  634.0  616.5  767.5  1120.0
```

Read counts data in from server

```
### -----
## Set up unix connection using sshfs
### -----
Sys.which('bash')
Sys.which('sh')
#echo hello world
#
system("sshfs stsmith@medusa.igs.umaryland.edu:/local/projects-t2/HRBV/
~/IGS/sshfs_medusa/", intern = FALSE,
       ignore.stdout = FALSE, ignore.stderr = FALSE,
       wait = TRUE, input = NULL, show.output.on.console = TRUE,
       minimized = FALSE, invisible = F)
```

```

manifest<-read.table("~/IGS/sshfs_medusa/Config/W100083533.manifest",sep =
"\t",header = T)
TRL_counts_table<-data.frame(Feature="None")

for(i in 1:length(manifest$Pre_QC_ID)){
  #i<-3
  sample_name<-as.character(manifest[i,"Pre_QC_ID"])
  sample_path<-
paste0("~/IGS/sshfs_medusa/TRL/alignment/TRL_",sample_name,"_aln_hg19/TRL_",s
ample_name,".hg_counts")
  new_table<-read.table(sample_path,header=T,sep="\t")
  names(new_table)<-c("Feature",sample_name)
  #sub_feature_set<-!new_table$Feature %in%
c(filtered_out_features,"mirna_info")
  ## Don't forget about ambiguous/none features!
  #new_table<-new_table[sub_feature_set,]
  TRL_counts_table<-merge(TRL_counts_table,new_table,by.x =
"Feature",by.y="Feature",all=T)
}

row.names(TRL_counts_table)<-TRL_counts_table$Feature
TRL_counts_table<-TRL_counts_table[!TRL_counts_table$Feature=="None",]
TRL_counts_table<-dplyr::select(TRL_counts_table,-
c(Feature,POSITIVE_CONTROL))
head(TRL_counts_table)
write.csv(TRL_counts_table,file=paste0(R_script_output_directory,TABLE_TRL_CO
UNTS_RAW),row.names = T,quote=F)

```

Create and Prepare Tables

```

TRL_counts_table<-
read.csv(file=paste0(R_script_output_directory,TABLE_TRL_COUNTS_RAW),row.name
s = 1)
write.csv(TRL_counts_table,file=paste0(thesis_tables_directory,TABLE_TRL_COUN
TS_RAW),row.names = T,quote=F)

## Filter out ambiguous, no feature reads. Make note of how many there are
TRL_counts_table[row.names(TRL_counts_table) %in%
c("alignment_not_unique","no_feature","ambiguous"),]

##
## VK2_GVAGINALIS_BCS_13HR_rep1
## alignment_not_unique 1659810
## ambiguous 72201
## no_feature 3075288
##
## VK2_LCRISPATUS_BCS_13HR_rep2
## alignment_not_unique 1412089
## ambiguous 36836
## no_feature 3279466
##
## VK2_LJENSENII_BCS_13HR_rep2 VK2_LINERS_BCS_13HR_rep2

```

## alignment_not_unique	1568882	657465
## ambiguous	43491	24472
## no_feature	3468681	2136668
## VK2_GVAGINALIS_BCS_13HR_rep2 VK2_MEDIA_BCS_13HR_rep1		
## alignment_not_unique	1019153	813706
## ambiguous	47530	32173
## no_feature	3914945	3257161
## VK2_LCRISPATUS_BCS_22HR_rep1		
## alignment_not_unique	888176	
## ambiguous	20079	
## no_feature	3186044	
## VK2_LJENSENII_BCS_22HR_rep1 VK2_LINERS_BCS_22HR_rep1		
## alignment_not_unique	1547911	479715
## ambiguous	47238	22808
## no_feature	2931450	2268805
## VK2_LJENSENII_BCS_4HR_rep2 VK2_LINERS_BCS_4HR_rep2		
## alignment_not_unique	1056023	1259736
## ambiguous	48041	43226
## no_feature	3037420	3488226
## VK2_GVAGINALIS_BCS_4HR_rep2 VK2_MEDIA_BCS_4HR_rep2		
## alignment_not_unique	712114	1230561
## ambiguous	26952	44876
## no_feature	2369104	3902295
## VK2_LCRISPATUS_BCS_4HR_rep1		
## alignment_not_unique	3511899	
## ambiguous	79107	
## no_feature	5717691	
## VK2_LJENSENII_BCS_4HR_rep1 VK2_LINERS_BCS_4HR_rep1		
## alignment_not_unique	673805	968908
## ambiguous	19770	33938
## no_feature	4664422	4209178
## VK2_GVAGINALIS_BCS_4HR_rep1 VK2_MEDIA_BCS_4HR_rep1		
## alignment_not_unique	1596106	1336499
## ambiguous	51607	52997
## no_feature	3658460	4357425
## VK2_LCRISPATUS_BCS_13HR_rep1		
## alignment_not_unique	1467776	
## ambiguous	41804	
## no_feature	2886596	
## VK2_LJENSENII_BCS_13HR_rep1 VK2_LINERS_BCS_13HR_rep1		
## alignment_not_unique	1009315	378146
## ambiguous	25940	11378
## no_feature	3342529	3070524
## VK2_MEDIA_BCS_22HR_rep1 VK2_LCRISPATUS_BCS_4HR_rep2		
## alignment_not_unique	1531300	774318
## ambiguous	66155	20755
## no_feature	5086258	3881756
## VK2_MEDIA_BCS_13HR_rep2 VK2_LCRISPATUS_BCS_22HR_rep2		
## alignment_not_unique	3011458	3026810
## ambiguous	92696	60297

```

## no_feature                3169404                4151315
## VK2_LJENSENII_BCS_22HR_rep2 VK2_LINERS_BCS_22HR_rep2
## alignment_not_unique      1190320                462056
## ambiguous                  30935                  14413
## no_feature                4117821                3010909
## VK2_GVAGINALIS_BCS_22HR_rep2
## alignment_not_unique      453069
## ambiguous                  16728
## no_feature                3168877
## VK2_GVAGINALIS_BCS_4HR_rep3 VK2_MEDIA_BCS_4HR_rep3
## alignment_not_unique      2380178                1313853
## ambiguous                  93189                  60327
## no_feature                4742894                4355267
## VK2_LCRISPATUS_BCS_13HR_rep3
## alignment_not_unique      188513
## ambiguous                  2133
## no_feature                2477532
## VK2_LJENSENII_BCS_13HR_rep3 VK2_LINERS_BCS_13HR_rep3
## alignment_not_unique      1389726                3545491
## ambiguous                  54050                  108724
## no_feature                4957907                4458395
## VK2_GVAGINALIS_BCS_13HR_rep3 VK2_MEDIA_BCS_13HR_rep3
## alignment_not_unique      5904365                2844329
## ambiguous                  230423                73038
## no_feature                5816169                5443374
## VK2_LCRISPATUS_BCS_22HR_rep3
## alignment_not_unique      7539963
## ambiguous                  171506
## no_feature                5283834
## VK2_LJENSENII_BCS_22HR_rep3 VK2_LINERS_BCS_22HR_rep3
## alignment_not_unique      6003165                2546317
## ambiguous                  195482                101360
## no_feature                5238683                4494104
## VK2_GVAGINALIS_BCS_22HR_rep3 VK2_MEDIA_BCS_22HR_rep3
## alignment_not_unique      2979388                2308331
## ambiguous                  87512                 104417
## no_feature                4729799                4141951

## Percentage of ambiguous/no feature/non unique reads & alignment stats
## (ambig_nofeat<-colSums(TRL_counts_table[row.names(TRL_counts_table) %in%
## c("alignment_not_unique","no_feature","ambiguous"),]))

## VK2_GVAGINALIS_BCS_13HR_rep1 VK2_LCRISPATUS_BCS_13HR_rep2
## 4807299                4728391
## VK2_LJENSENII_BCS_13HR_rep2 VK2_LINERS_BCS_13HR_rep2
## 5081054                2818605
## VK2_GVAGINALIS_BCS_13HR_rep2 VK2_MEDIA_BCS_13HR_rep1
## 4981628                4103040
## VK2_LCRISPATUS_BCS_22HR_rep1 VK2_LJENSENII_BCS_22HR_rep1
## 4094299                4526599

```

```

##      VK2_LINERS_BCS_22HR_rep1      VK2_LJENSENII_BCS_4HR_rep2
##                                2771328                                4141484
##      VK2_LINERS_BCS_4HR_rep2      VK2_GVAGINALIS_BCS_4HR_rep2
##                                4791188                                3108170
##      VK2_MEDIA_BCS_4HR_rep2      VK2_LCRISPATUS_BCS_4HR_rep1
##                                5177732                                9308697
##      VK2_LJENSENII_BCS_4HR_rep1      VK2_LINERS_BCS_4HR_rep1
##                                5357997                                5212024
##      VK2_GVAGINALIS_BCS_4HR_rep1      VK2_MEDIA_BCS_4HR_rep1
##                                5306173                                5746921
##      VK2_LCRISPATUS_BCS_13HR_rep1      VK2_LJENSENII_BCS_13HR_rep1
##                                4396176                                4377784
##      VK2_LINERS_BCS_13HR_rep1      VK2_MEDIA_BCS_22HR_rep1
##                                3460048                                6683713
##      VK2_LCRISPATUS_BCS_4HR_rep2      VK2_MEDIA_BCS_13HR_rep2
##                                4676829                                6273558
##      VK2_LCRISPATUS_BCS_22HR_rep2      VK2_LJENSENII_BCS_22HR_rep2
##                                7238422                                5339076
##      VK2_LINERS_BCS_22HR_rep2      VK2_GVAGINALIS_BCS_22HR_rep2
##                                3487378                                3638674
##      VK2_GVAGINALIS_BCS_4HR_rep3      VK2_MEDIA_BCS_4HR_rep3
##                                7216261                                5729447
##      VK2_LCRISPATUS_BCS_13HR_rep3      VK2_LJENSENII_BCS_13HR_rep3
##                                2668178                                6401683
##      VK2_LINERS_BCS_13HR_rep3      VK2_GVAGINALIS_BCS_13HR_rep3
##                                8112610                                11950957
##      VK2_MEDIA_BCS_13HR_rep3      VK2_LCRISPATUS_BCS_22HR_rep3
##                                8360741                                12995303
##      VK2_LJENSENII_BCS_22HR_rep3      VK2_LINERS_BCS_22HR_rep3
##                                11437330                                7141781
##      VK2_GVAGINALIS_BCS_22HR_rep3      VK2_MEDIA_BCS_22HR_rep3
##                                7796699                                6554699

```

```

(aligned<-colSums(TRL_counts_table[!row.names(TRL_counts_table) %in%
c("alignment_not_unique","no_feature","ambiguous"),]))

```

```

##      VK2_GVAGINALIS_BCS_13HR_rep1      VK2_LCRISPATUS_BCS_13HR_rep2
##                                3483930                                3024389
##      VK2_LJENSENII_BCS_13HR_rep2      VK2_LINERS_BCS_13HR_rep2
##                                2974822                                1729117
##      VK2_GVAGINALIS_BCS_13HR_rep2      VK2_MEDIA_BCS_13HR_rep1
##                                3155228                                2030629
##      VK2_LCRISPATUS_BCS_22HR_rep1      VK2_LJENSENII_BCS_22HR_rep1
##                                1416161                                2950478
##      VK2_LINERS_BCS_22HR_rep1      VK2_LJENSENII_BCS_4HR_rep2
##                                1532055                                3206459
##      VK2_LINERS_BCS_4HR_rep2      VK2_GVAGINALIS_BCS_4HR_rep2
##                                3115872                                1880874
##      VK2_MEDIA_BCS_4HR_rep2      VK2_LCRISPATUS_BCS_4HR_rep1
##                                3207936                                5558855

```

```

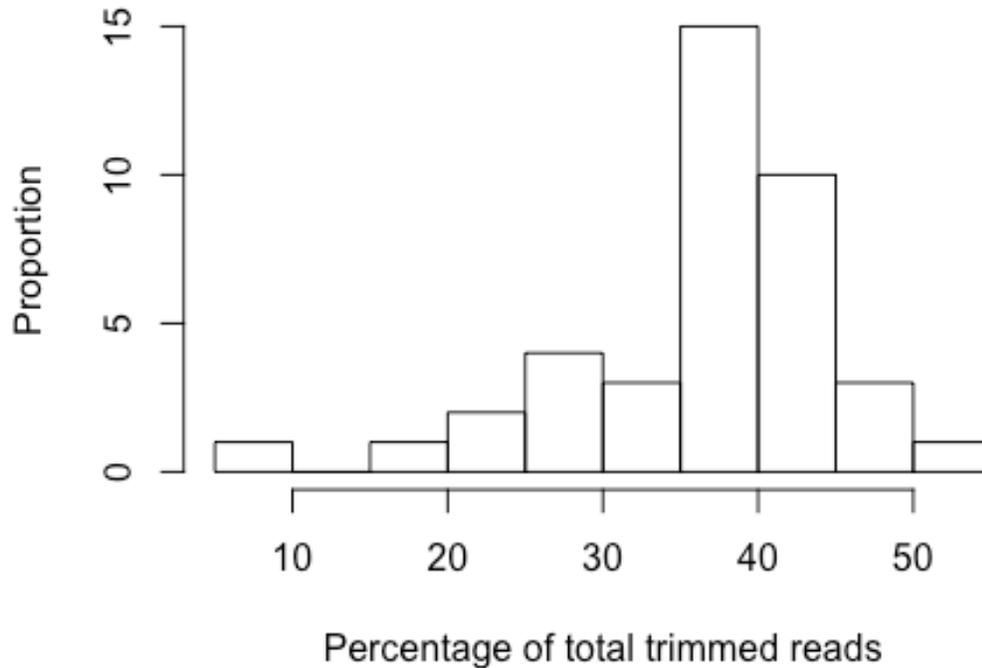
## VK2_LJENSENII_BCS_4HR_rep1 VK2_LINERS_BCS_4HR_rep1
## 1533479 2645213
## VK2_GVAGINALIS_BCS_4HR_rep1 VK2_MEDIA_BCS_4HR_rep1
## 3555610 3745190
## VK2_LCRISPATUS_BCS_13HR_rep1 VK2_LJENSENII_BCS_13HR_rep1
## 3386120 1921098
## VK2_LINERS_BCS_13HR_rep1 VK2_MEDIA_BCS_22HR_rep1
## 863055 4670961
## VK2_LCRISPATUS_BCS_4HR_rep2 VK2_MEDIA_BCS_13HR_rep2
## 1648935 4892080
## VK2_LCRISPATUS_BCS_22HR_rep2 VK2_LJENSENII_BCS_22HR_rep2
## 4392899 2049302
## VK2_LINERS_BCS_22HR_rep2 VK2_GVAGINALIS_BCS_22HR_rep2
## 1021441 1354926
## VK2_GVAGINALIS_BCS_4HR_rep3 VK2_MEDIA_BCS_4HR_rep3
## 5603541 5132964
## VK2_LCRISPATUS_BCS_13HR_rep3 VK2_LJENSENII_BCS_13HR_rep3
## 161415 3802258
## VK2_LINERS_BCS_13HR_rep3 VK2_GVAGINALIS_BCS_13HR_rep3
## 6838588 10538780
## VK2_MEDIA_BCS_13HR_rep3 VK2_LCRISPATUS_BCS_22HR_rep3
## 4873897 8448815
## VK2_LJENSENII_BCS_22HR_rep3 VK2_LINERS_BCS_22HR_rep3
## 8633553 5445716
## VK2_GVAGINALIS_BCS_22HR_rep3 VK2_MEDIA_BCS_22HR_rep3
## 5545096 7577938

alignment_stats<-
data.frame(aligned=aligned,non_aligned=ambig_nofeat,aligned.percent=100*aligned/colSums(TRL_counts_table))
write.csv(alignment_stats,file =
paste0(thesis_tables_directory,TABLE_TRL_ALIGNSTATS))

## Look at proportion of Ambiguous/no feature/non-unique
hist(alignment_stats$aligned.percent,main="Percentage of Aligned
reads",xlab="Percentage of total trimmed reads",ylab="Proportion")

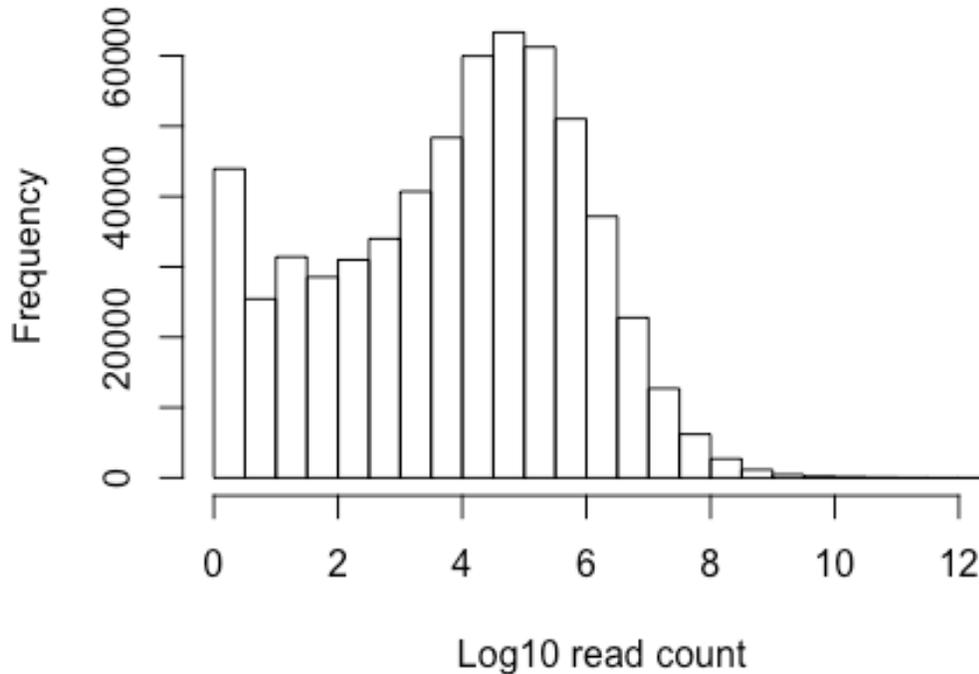
```

Percentage of Aligned reads



```
ambig_nofeat_readpercent.high<-alignment_stats[(100-  
alignment_stats$aligned.percent)>80,]  
paste0("The following samples have ambiguous reads >80%: ",  
str_c(row.names(ambig_nofeat_readpercent.high),collapse = ", "))  
  
## [1] "The following samples have ambiguous reads >80%:  
VK2_LINERS_BCS_13HR_rep1, VK2_LCRISPATUS_BCS_13HR_rep3"  
  
#Filter out ambiguous reads from counts  
TRL_counts_table<-TRL_counts_table[!row.names(TRL_counts_table) %in%  
c("alignment_not_unique", "no_feature", "ambiguous"),]  
TRL_counts_table<-TRL_counts_table[,order(names(TRL_counts_table))]  
hist(log(as.matrix(TRL_counts_table)),main="log10 read counts, all  
samples",xlab="Log10 read count")
```

log10 read counts, all samples



```
##Create Design Table
TRL_design<-data.frame(Sample=names(TRL_counts_table))
row.names(TRL_design)<-TRL_design$Sample
TRL_design<-separate(TRL_design,Sample,sep = "_",into =
c("CellLine","BCS","DROP","ExposureTime","rep")) %>% dplyr::select(-
c(CellLine,DROP))

## Create replicate #, lactobacillus indicator. Then combine BCS and exposure
time for 'group1' for eventually creating constrasts
TRL_design$rep<-as.numeric(gsub(TRL_design$rep,pattern = "rep",replacement =
""))
TRL_design$Lactobacillus<-grepl(TRL_design$BCS,pattern = "^L")
TRL_design$group1 <-
factor(paste(TRL_design$BCS,TRL_design$ExposureTime,sep="."))
TRL_design$BCS<-factor(TRL_design$BCS,levels =
c("LCRISPATUS","LJENSENII","LINERS","GVAGINALIS","MEDIA"),ordered = T)
TRL_design$ExposureTime<-factor(TRL_design$ExposureTime,levels =
c("4HR","13HR","22HR"),ordered=T)
sample_order<-with(TRL_design,order(ExposureTime,BCS,rep))

##Make an expression set object for coupled handling of counts and design
matrix
TRL_design<-TRL_design[sample_order,]
```

```

TRL_counts_table<-TRL_counts_table[,sample_order]
TRL_counts_meta<-ExpressionSet(assayData =
as.matrix(TRL_counts_table),phenoData = AnnotatedDataFrame(TRL_design))

```

Plot Replicates

```

## This may take a while given there are ~20,000 points to plot for each
comparison. Consider trimming the lower expressed reads by using rmlow =
log(10,10) for example
postscript(paste0(R_script_output_directory,"TRL_RNASeq_Replicate_plots.eps")
,width = 10,height = 8)
hist(colSums(exprs(TRL_counts_meta)))

plot_replicates(TRL_counts_meta,BCS.selection =
c("LCRISPATUS"),ExposureTime.selection = c("4HR","13HR","22HR"),logt=T)
plot_replicates(TRL_counts_meta,BCS.selection =
c("LJENSENII"),ExposureTime.selection = c("4HR","13HR","22HR"),logt=T)
plot_replicates(TRL_counts_meta,BCS.selection =
c("LINERS"),ExposureTime.selection = c("4HR","13HR","22HR"),logt=T)
plot_replicates(TRL_counts_meta,BCS.selection =
c("GVAGINALIS"),ExposureTime.selection = c("4HR","13HR","22HR"),logt=T)
plot_replicates(TRL_counts_meta,BCS.selection =
c("MEDIA"),ExposureTime.selection = c("4HR","13HR","22HR"),logt=T)

dev.off()

```

Drop samples

```

dropped_TRL_samples<-
c("VK2_MEDIA_BCS_4HR_rep3","VK2_LCRISPATUS_BCS_13HR_rep3") ## determined from
replicate plots and # of ambiguous (rRNA) reads.

##Drop poor QC samples
TRL_counts_meta.qc<-subset_ExpressionSet(TRL_counts_meta,filterOut
=dropped_TRL_samples )

## Proportion of genes with at least 1 read across all samples
sum(rowSums(exprs(TRL_counts_meta.qc)>0)>0)/nrow(exprs(TRL_counts_meta.qc))

## [1] 0.7899965

#exprs(TRL_counts_meta.qc)[rowSums(exprs(TRL_counts_meta.qc)>0)>0,]

##Summary of # of samples with at least one read across all samples
summary(rowSums(exprs(TRL_counts_meta.qc)[rowSums(exprs(TRL_counts_meta.qc)>0
)>0,]>0))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.0   19.0   38.0   28.8   38.0   38.0

##Total post QC samples:
ncol(exprs(TRL_counts_meta.qc))

```

```
## [1] 38

## Summary of remaining total read counts
summary(colSums(exprs(TRL_counts_meta.qc)))

##      Min.  1st Qu.  Median    Mean 3rd Qu.    Max.
## 863100 1948000 3207000 3796000 4888000 10540000

(reps_per_treatment<-
ddply(data.frame(pData(TRL_counts_meta.qc)),c("BCS","ExposureTime"),summarise
,n=length(BCS)))

##           BCS ExposureTime n
## 1  LCRISPATUS           4HR 2
## 2  LCRISPATUS          13HR 2
## 3  LCRISPATUS          22HR 3
## 4   LJENSENII           4HR 2
## 5   LJENSENII          13HR 3
## 6   LJENSENII          22HR 3
## 7     LINERS           4HR 2
## 8     LINERS          13HR 3
## 9     LINERS          22HR 3
## 10  GVAGINALIS           4HR 3
## 11  GVAGINALIS          13HR 3
## 12  GVAGINALIS          22HR 2
## 13     MEDIA           4HR 2
## 14     MEDIA          13HR 3
## 15     MEDIA          22HR 2
```

Create model matrix

This is done outside edgeR GLM as it is used by other chunks, but edgeR chunk is not evaluated

```
## take counts table design created above and make edgeR object
design <- model.matrix(~0+group1,data = pData(TRL_counts_meta.qc))
```

edgeR GLM FIT

Not evualted as it takes some time. Change to eval=T to re-compute edgeR results

```
## Estimate dispersion, calc norm facots, and fit to GLM
y<-DGEList(exprs(TRL_counts_meta.qc))
y<-estimateDisp(y,design)
y<-calcNormFactors(y)
plotBCV(y,main="BCV Plot")
TRL_glmFit <- glmFit(y, design)

save(TRL_glmFit,file = paste0(R_script_output_directory,"TRL_glmFit.RData"))
```

Perform edgeR LRT

```
load(paste0(R_script_output_directory,"TRL_glmFit.RData"))
## Go through each pair-wise comparison in contrasts and compute
differential expression using glmLRT
contr<-makeContrasts(
  LCGV.4="group1LCRISPATUS.4HR-group1GVAGINALIS.4HR",
  LCM.4="group1LCRISPATUS.4HR-group1MEDIA.4HR",
  LJGV.4="group1LJENSENII.4HR-group1GVAGINALIS.4HR",
  LJM.4="group1LJENSENII.4HR-group1MEDIA.4HR",
  LIGV.4="group1LINERS.4HR-group1GVAGINALIS.4HR",
  LIM.4="group1LINERS.4HR-group1MEDIA.4HR",
  LCLJ.4="group1LCRISPATUS.4HR-group1LJENSENII.4HR",
  LCLI.4="group1LCRISPATUS.4HR-group1LINERS.4HR",
  LJLI.4="group1LJENSENII.4HR-group1LINERS.4HR",
  GVM.4="group1GVAGINALIS.4HR-group1MEDIA.4HR",

  LCGV.13="group1LCRISPATUS.13HR-group1GVAGINALIS.13HR",
  LCM.13="group1LCRISPATUS.13HR-group1MEDIA.13HR",
  LJGV.13="group1LJENSENII.13HR-group1GVAGINALIS.13HR",
  LJM.13="group1LJENSENII.13HR-group1MEDIA.13HR",
  LIGV.13="group1LINERS.13HR-group1GVAGINALIS.13HR",
  LIM.13="group1LINERS.13HR-group1MEDIA.13HR",
  LCLJ.13="group1LCRISPATUS.13HR-group1LJENSENII.13HR",
  LCLI.13="group1LCRISPATUS.13HR-group1LINERS.13HR",
  LJLI.13="group1LJENSENII.13HR-group1LINERS.13HR",
  GVM.13="group1GVAGINALIS.13HR-group1MEDIA.13HR",

  LCGV.22="group1LCRISPATUS.22HR-group1GVAGINALIS.22HR",
  LCM.22="group1LCRISPATUS.22HR-group1MEDIA.22HR",
  LJGV.22="group1LJENSENII.22HR-group1GVAGINALIS.22HR",
  LJM.22="group1LJENSENII.22HR-group1MEDIA.22HR",
  LIGV.22="group1LINERS.22HR-group1GVAGINALIS.22HR",
  LIM.22="group1LINERS.22HR-group1MEDIA.22HR",
  LCLJ.22="group1LCRISPATUS.22HR-group1LJENSENII.22HR",
  LCLI.22="group1LCRISPATUS.22HR-group1LINERS.22HR",
  LJLI.22="group1LJENSENII.22HR-group1LINERS.22HR",
  GVM.22="group1GVAGINALIS.22HR-group1MEDIA.22HR",

  levels=design
)

desets<-list()
comparisons<-names(data.frame(contr))

#postscript(paste0(R_script_output_directory,"TRL_SmearPlots.eps"),height =
8,width = 10)

for(comparison in comparisons){
  #comparison<-"LCGV.13"
```

```

comp<-glmLRT(TRL_glmFit,contrast = contr[,comparison]) ## DE using
constrast

de.table<-comp$table[abs(comp$table$logFC)>1 & comp$table$logCPM>1 &
p.adjust(comp$table$PValue,method = "fdr")<=0.01,]
comp$table$PValue.adj<-p.adjust(comp$table$PValue,method = "fdr")
desets[[comparison]]<-list(detags=nrow(de.table),fulltable=comp$table)
#plotSmear(comp,de.tags = names(de.table),main=paste0(comparison,"
",length(detags)," DE genes"))
}

```

Upload DE table to IPA, run IPA pathway analysis, then save the pathway comparisons to file Performed March-April 2017. See thesis for version/db builds.

Read in IPA Results

```

##Timecourse files are from IPA- contain pathway list and activation z
values.

pathway_zscore_files<-list.files(path = R_script_input_directory, pattern =
"timecourse.txt")
pathway_zscores<-data.frame(Canonical.Pathway="DROP")
functions<-
data.frame(Comparison="DROP",Categories="DROP",Diseases.or.Functions.Annotati
on="DROP",p.Value=1,Predicted.Activation.State="DROP",Activation.z.score=0,Fl
ags="DROP", Molecules="DROP")

for(tab in pathway_zscore_files){
#tab<- "LCGV_LCM_timecourse.txt"
newt<- read.table(paste0(R_script_input_directory,tab),header = T,sep =
"\t",skip = 1,na.strings = "N/A",stringsAsFactors = F) %>% dplyr::select(-X)
pathway_zscores<-merge(newt,pathway_zscores,all = T)
}

## Put pathway z scores into matrices and then split absed on G. vag or
medium reference
pathway_zscores.matrix<-as.matrix(pathway_zscores[,2:ncol(pathway_zscores)])
row.names(pathway_zscores.matrix)<-pathway_zscores$Canonical.Pathway
pathway_zscores.matrix<-
pathway_zscores.matrix[rowSums(abs(pathway_zscores.matrix),na.rm =
T)>0,colSums(abs(pathway_zscores.matrix),na.rm = T)>0]

pathway_zscores.melt<-melt(pathway_zscores,id.vars = "Canonical.Pathway")
row.names(pathway_zscores)<-pathway_zscores$Canonical.Pathway
pathway_zscores.matrix[is.na(pathway_zscores.matrix)]<-0

## Create a design table for the pathways
Pathways_design<-
data.frame(comparison=names(data.frame(pathway_zscores.matrix)))

```

```

Pathways_design$comp<-
sapply(strsplit(as.character(Pathways_design$comparison), "\\."), function(x)
x[[1]])
Pathways_design$ExposureTime<-
sapply(strsplit(as.character(Pathways_design$comparison), "\\."), function(x)
x[[2]])
Pathways_design$L<-sapply(strsplit(Pathways_design$comp, split =
"*"), function(x) paste0(x[[1]],x[[2]]))
Pathways_design$ref<-sapply(strsplit(Pathways_design$comp, split =
"*"), function(x) paste0(x[[3]]))
row.names(Pathways_design)<-Pathways_design$comparison

## Map mapthway names to classifications
path_type_map<-
read.csv(paste0(R_script_input_directory, "Pathway_classification.csv"), string
sAsFactors = F)

## Subset to look at only cell culture medium references
medium_comparisons<-
names(data.frame(pathway_zscores.matrix))[grep1(names(data.frame(pathway_zsco
res.matrix)), pattern = "M")]
pathway_zscores.matrix.medium<-
pathway_zscores.matrix[, names(data.frame(pathway_zscores.matrix)) %in%
medium_comparisons]

## Summary table for pathway classification
summary_pathways<-
cbind(data.frame(num_cycle.p=colSums(pathway_zscores.matrix.medium[row.names(
data.frame(pathway_zscores.matrix.medium)) %in%
path_type_map[path_type_map$class=="c", "pathway"], ]>2)),

data.frame(num_cycle.n=colSums(pathway_zscores.matrix.medium[row.names(data.f
rame(pathway_zscores.matrix.medium)) %in%
path_type_map[path_type_map$class=="c", "pathway"], ]<(-2))),

data.frame(num_immune.p=colSums(pathway_zscores.matrix.medium[row.names(data.
frame(pathway_zscores.matrix.medium)) %in%
path_type_map[path_type_map$class=="i", "pathway"], ]>2)),

data.frame(num_immune.n=colSums(pathway_zscores.matrix.medium[row.names(data.
frame(pathway_zscores.matrix.medium)) %in%
path_type_map[path_type_map$class=="i", "pathway"], ]<(-2))),

data.frame(num_immune_pro.n=colSums(pathway_zscores.matrix.medium[row.names(d
ata.frame(pathway_zscores.matrix.medium)) %in%
path_type_map[path_type_map$class=="pro", "pathway"], ]<(-2))),

```

```

data.frame(num_immune_pro.p=colSums(pathway_zscores.matrix.medium[row.names(d
ata.frame(pathway_zscores.matrix.medium)) %in%
path_type_map[path_type_map$X=="pro", "pathway"], ]>2)))

summary_pathways$comparison<-row.names(summary_pathways)
summary_pathways<-merge(summary_pathways, Pathways_design)

## Write table summarizing the number of pathways above or below z score,
grouped by pathway category
write.csv(summary_pathways, paste0(thesis_tables_directory, TABLE_TRL_SUMMARY_P
ATHWAYS), row.names=F, quote=F)

write.csv(pathway_zscores.matrix.medium, paste0(thesis_tables_directory, TABLE_
TRL_PATHWAY_Z_SCORES), row.names=T, quote=F)

##Sort # of z>2 pathways by negative cycle, positive immune and negative
immune
summary_pathways[order(summary_pathways$num_cycle.n, decreasing = T),]

##      comparison num_cycle.p num_cycle.n num_immune.p num_immune.n
## 4      LCM.13          1          4          2          0
## 5      LCM.22          0          3          0          1
## 6      LCM.4           3          2          11         0
## 10     LJM.13          0          2          1          1
## 11     LJM.22          0          2          1          1
## 3      GVM.4           3          1          18         0
## 8      LIM.22          0          1          3          0
## 12     LJM.4           1          1          2          0
## 1      GVM.13          1          0          8          0
## 2      GVM.22          0          0          2          0
## 7      LIM.13          0          0          7          0
## 9      LIM.4           4          0          16         0
##      num_immune_pro.n num_immune_pro.p comp ExposureTime L ref
## 4          0          1 LCM          13 LC M
## 5          0          0 LCM          22 LC M
## 6          0          5 LCM           4 LC M
## 10         0          1 LJM          13 LJ M
## 11         0          1 LJM          22 LJ M
## 3          0          8 GVM           4 GV M
## 8          0          1 LIM          22 LI M
## 12         0          2 LJM           4 LJ M
## 1          0          3 GVM          13 GV M
## 2          0          1 GVM          22 GV M
## 7          0          4 LIM          13 LI M
## 9          0          7 LIM           4 LI M

summary_pathways[order(summary_pathways$num_immune.p, decreasing = T),]

```

```

##      comparison num_cycle.p num_cycle.n num_immune.p num_immune.n
## 3      GVM.4      3      1      18      0
## 9      LIM.4      4      0      16      0
## 6      LCM.4      3      2      11      0
## 1      GVM.13     1      0      8      0
## 7      LIM.13     0      0      7      0
## 8      LIM.22     0      1      3      0
## 2      GVM.22     0      0      2      0
## 4      LCM.13     1      4      2      0
## 12     LJM.4      1      1      2      0
## 10     LJM.13     0      2      1      1
## 11     LJM.22     0      2      1      1
## 5      LCM.22     0      3      0      1
##      num_immune_pro.n num_immune_pro.p comp ExposureTime L ref
## 3      0      8 GVM      4 GV M
## 9      0      7 LIM      4 LI M
## 6      0      5 LCM      4 LC M
## 1      0      3 GVM     13 GV M
## 7      0      4 LIM     13 LI M
## 8      0      1 LIM     22 LI M
## 2      0      1 GVM     22 GV M
## 4      0      1 LCM     13 LC M
## 12     0      2 LJM      4 LJ M
## 10     0      1 LJM     13 LJ M
## 11     0      1 LJM     22 LJ M
## 5      0      0 LCM     22 LC M

summary_pathways[order(summary_pathways$num_immune.n,decreasing = T),]

##      comparison num_cycle.p num_cycle.n num_immune.p num_immune.n
## 5      LCM.22     0      3      0      1
## 10     LJM.13     0      2      1      1
## 11     LJM.22     0      2      1      1
## 1      GVM.13     1      0      8      0
## 2      GVM.22     0      0      2      0
## 3      GVM.4      3      1      18     0
## 4      LCM.13     1      4      2      0
## 6      LCM.4      3      2      11     0
## 7      LIM.13     0      0      7      0
## 8      LIM.22     0      1      3      0
## 9      LIM.4      4      0      16     0
## 12     LJM.4      1      1      2      0
##      num_immune_pro.n num_immune_pro.p comp ExposureTime L ref
## 5      0      0 LCM     22 LC M
## 10     0      1 LJM     13 LJ M
## 11     0      1 LJM     22 LJ M
## 1      0      3 GVM     13 GV M
## 2      0      1 GVM     22 GV M
## 3      0      8 GVM      4 GV M
## 4      0      1 LCM     13 LC M

```

```

## 6          0          5 LCM          4 LC   M
## 7          0          4 LIM          13 LI   M
## 8          0          1 LIM          22 LI   M
## 9          0          7 LIM          4 LI   M
## 12         0          2 LJM          4 LJ   M

## Use pathways that are expressed abs(z-score)>2 in 10% of comparisons
pathway_zscores.matrix.medium<-
pathway_zscores.matrix.medium[(rowSums(abs(pathway_zscores.matrix.medium)>2))
>=.1*ncol(pathway_zscores.matrix.medium),]

## Clean up some of the pathway names in pathway map
path_type_map[path_type_map$pathway=="NF-_B Signaling","pathway"]<-"NF-κB
Signaling"
path_type_map[path_type_map$pathway=="PKC_ Signaling in T
Lymphocytes","pathway"]<-"PKCθ Signaling in T Lymphocytes"

path_type_map[path_type_map$X=="pro","class"]<-"pro"

## Add in pathway classification based on mapping file
pathway_zscores.matrix.medium<-data.frame(pathway_zscores.matrix.medium)
pathway_zscores.matrix.medium$pathway<-
row.names(pathway_zscores.matrix.medium)
pathway_zscores.matrix.medium<-
merge(pathway_zscores.matrix.medium,path_type_map) ## maps classifications to
pathway names
row.names(pathway_zscores.matrix.medium)<-
pathway_zscores.matrix.medium$pathway

table(path_type_map[path_type_map$pathway %in%
row.names(data.frame(pathway_zscores.matrix.medium)),"class"])

##
##   c   i   o pro
##  11  12   5   7

## Melt the pathway z-scores table
pathway_zscores.matrix.medium.melt<-melt(pathway_zscores.matrix.medium,id.vars =
c("class","X","pathway"))
pathway_zscores.matrix.medium.melt<-
dplyr::rename(pathway_zscores.matrix.medium.melt,"comparison"=variable)
pathway_zscores.matrix.medium.melt<-
merge(pathway_zscores.matrix.medium.melt,Pathways_design)

## Ensure the exposure times and BCS are ordered
summary_pathways$ExposureTime<-as.numeric(summary_pathways$ExposureTime)
summary_pathways$ExposureTime<-factor(summary_pathways$ExposureTime,levels =
c(4,13,22),ordered = T)
pathway_zscores.matrix.medium.melt$comp<-
factor(pathway_zscores.matrix.medium.melt$comp,levels =

```

```

c("LCM","LJM","LIM","GVM"),ordered = T)
pathway_zscores.medium.melt$ExposureTime<-
factor(pathway_zscores.medium.melt$ExposureTime,levels =
c("4","13","22"),ordered = T)
pathway_zscores.medium.melt$x<-
paste(pathway_zscores.medium.melt$comp,pathway_zscores.medium.melt$pathway)

##Clean up the pathway names for better plotting
pathway_zscores.medium.melt[pathway_zscores.medium.melt$pathway=="Role of IL-
17F in Allergic Inflammatory Airway Diseases","pathway"]<-"IL-17F in Allgc.
Inflam. Arwy Dis."
pathway_zscores.medium.melt[pathway_zscores.medium.melt$pathway=="Role of
Pattern Recognition Receptors in Recognition of Bacteria and
Viruses","pathway"]<-"PRRs/ Bacteria and Viruses"
pathway_zscores.medium.melt[pathway_zscores.medium.melt$pathway=="Production
of Nitric Oxide and Reactive Oxygen Species in Macrophages","pathway"]<-
"Production of NO and ROS in Macrophages"
pathway_zscores.medium.melt[pathway_zscores.medium.melt$pathway=="PKCθ
Signaling in T Lymphocytes","pathway"]<-"PKCθ Signaling"
pathway_zscores.medium.melt[pathway_zscores.medium.melt$pathway=="PI3K
Signaling in B Lymphocytes","pathway"]<-"PI3K Signaling"

##Subset z-scores table by immune (proinflammatory) pathways or cell cycle
pathways
pathway_zscores.medium.melt.immune<-
dplyr::filter(pathway_zscores.medium.melt,class %in% c("pro", "i"))
pathway_zscores.medium.melt.cycle<-
dplyr::filter(pathway_zscores.medium.melt,class %in% c("c"))

##Write figures
postscript(paste0(thesis_figures_directory,FIGURE_COMBINED_PATHWAYS_IMMUNE,".
ps"),width=10,height=8)

unique(dplyr::select(pathway_zscores.medium.melt,c(class,pathway)))

##      class                pathway
## 1    pro      Acute Phase Response Signaling
## 2      c      Antioxidant Action of Vitamin C
## 3      c                      ATM Signaling
## 4    pro      B Cell Receptor Signaling
## 5      c Cholecystokinin/Gastrin-mediated Signaling
## 6      c      Colorectal Cancer Metastasis Signaling
## 7      c      Cyclins and Cell Cycle Regulation
## 8      c      Death Receptor Signaling
## 9      i      Dendritic Cell Maturation
## 10     c      Estrogen-mediated S-phase Entry
## 11    pro      HMGB1 Signaling
## 12    pro      IL-1 Signaling
## 13     i      IL-6 Signaling
## 14     i      IL-8 Signaling

```

```

## 15      c          ILK Signaling
## 16      i          iNOS Signaling
## 17      i          Interferon Signaling
## 18      o          LXR/RXR Activation
## 19     pro          MIF Regulation of Innate Immunity
## 20      c          Mitotic Roles of Polo-Like Kinase
## 21      i          NF-κB Signaling
## 22      o          NRF2-mediated Oxidative Stress Response
## 23      i          p38 MAPK Signaling
## 24      i          PI3K Signaling
## 25      i          PKCθ Signaling
## 26      o          PPAR Signaling
## 27      i          Production of NO and ROS in Macrophages
## 28      c          RANK Signaling in Osteoclasts
## 29      c          Role of BRCA1 in DNA Damage Response
## 30      i          IL-17F in Allgc. Inflamm. Arwy Dis.
## 31     pro          PRRs/ Bacteria and Viruses
## 32     pro          Toll-like Receptor Signaling
## 33      i          TREM1 Signaling
## 34      o          Type I Diabetes Mellitus Signaling
## 35      o          UVA-Induced MAPK Signaling

paste0("Number of pathways belonging to each class:")

## [1] "Number of pathways belonging to each class:"

table(unique(dplyr::select(pathway_zscores.medium.melt,c(class,pathway))) %>%
dplyr::select(class))

##
##   c   i   o pro
## 11 12   5   7

paste0(c("the remaining 5 pathways did not belong to either cell cycle or
immunity:",str_c(unique(dplyr::select(pathway_zscores.medium.melt,c(class,pat
hway)))) %>% dplyr::filter(class=="o") %>% dplyr::select(pathway),collapse =
", ")))

## [1] "the remaining 5 pathways did not belong to either cell cycle or
immunity:"
## [2] "c(\\"LXR/RXR Activation\\", \\"NRF2-mediated Oxidative Stress
Response\\", \\"PPAR Signaling\\", \\"Type I Diabetes Mellitus Signaling\\",
\\"UVA-Induced MAPK Signaling\\")"

ggplot(pathway_zscores.medium.melt.immune)+geom_tile(aes(x=ExposureTime,y=pat
hway,fill=value))+
scale_fill_gradient2(high="red",mid="white",low="blue",
na.value="yellow", midpoint=0)+facet_wrap(~comp,nrow=1)+
#mBio+
theme_bw() + theme(text = element_text(colour = "black",size=12))
#theme(text = element_text(size=12),plot.margin = unit(c(0,30,0,30),units =

```

```

"pt"))

dev.off()

## quartz_off_screen
##                2

cairo_ps(paste0(thesis_figures_directory,FIGURE_COMBINED_PATHWAYS_CYCLE,".eps
"),width=10,height=8)

ggplot(pathway_zscores.medium.melt.cycle)+geom_tile(aes(x=ExposureTime,y=path
way,fill=value))+
scale_fill_gradient2(high="red",mid="white",low="blue",
na.value="yellow", midpoint=0)+facet_wrap(~comp,nrow=1)+
  mBio+
  theme(text = element_text(size=12),plot.margin = unit(c(0,30,0,0),units =
"pt"))

dev.off()

## quartz_off_screen
##                2

```

Extract and plot logFC values from LRT table

```

## Pull out DE tables from each comparison within list, and write this table
to file
edgeR_results<-data.frame(gene=row.names(exprs(TRL_counts_meta.qc)))
row.names(edgeR_results)<-edgeR_results$gene

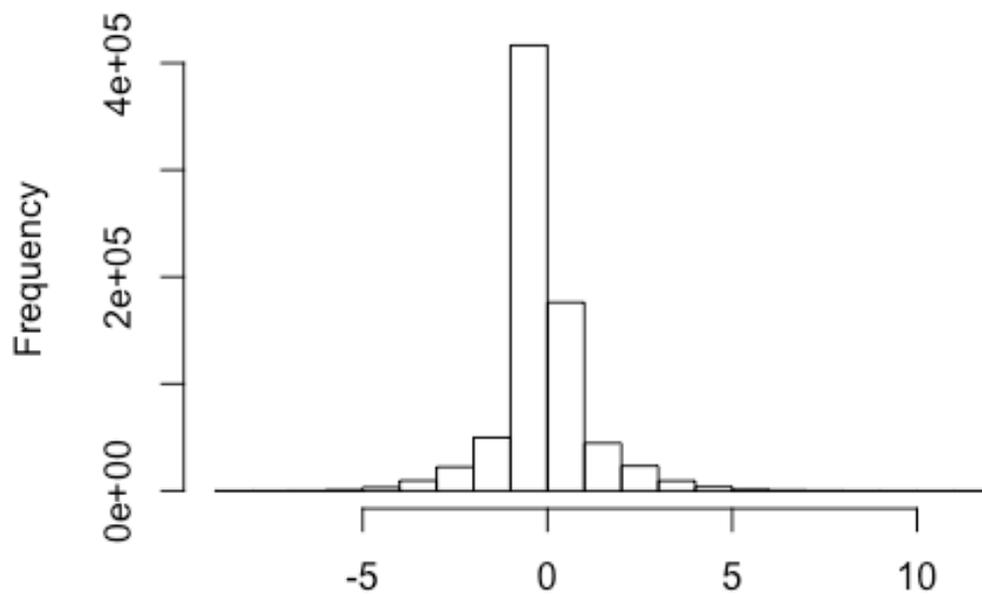
num_de_genes<-data.frame(num_de_genes=sapply(desets,function(x) x[[1]]))
num_de_genes$comparison<-row.names(num_de_genes)
write.csv(num_de_genes[num_de_genes$comparison %in%
medium_comparisons,],file=paste0(m=thesis_tables_directory,TABLE_TRL_NUMDEGEN
ES))

##Extract genes from the list
for(i in names(desets)){
  #i<-"LCGV.4"
  tmp.df<-data.frame(desets[[i]]$fulltable)
  names(tmp.df)<-paste0(i,".",names(tmp.df))
  tmp.df$gene<-row.names(desets[[i]]$fulltable)
  edgeR_results<-join(edgeR_results,tmp.df,"gene")
}

hist(as.matrix(dplyr::select(edgeR_results,ends_with("logFC"))))

```

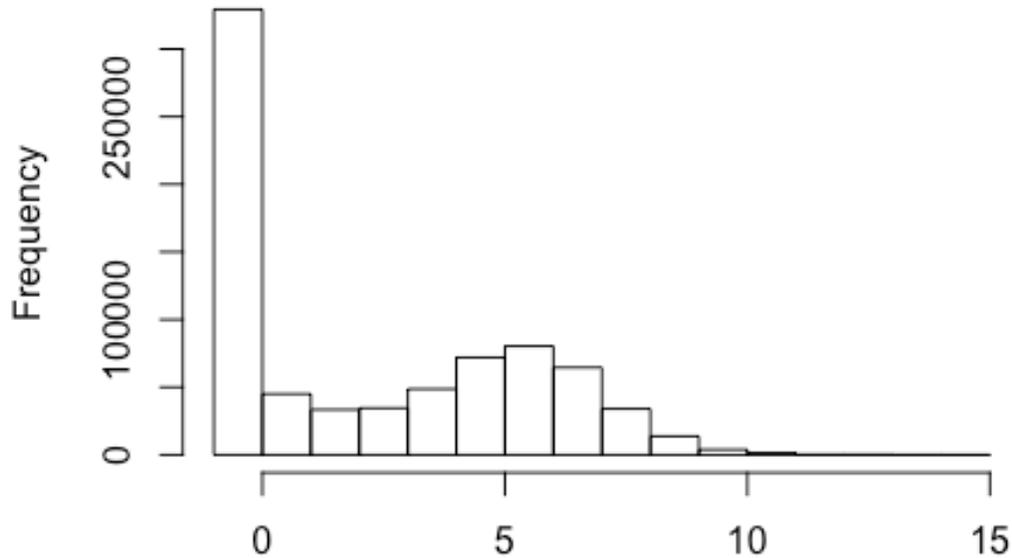
m of `as.matrix(dplyr::select(edgeR_results, ends_wi`



`as.matrix(dplyr::select(edgeR_results, ends_with("logFC")))`

`hist(as.matrix(dplyr::select(edgeR_results, ends_with("logCPM"))))`

n of as.matrix(dplyr::select(edgeR_results, ends_witl



as.matrix(dplyr::select(edgeR_results, ends_with("logCPM")))

```
quantile(as.matrix(dplyr::select(edgeR_results, ends_with("logFC"))), probs =  
c(0.68, 0.95))
```

```
##           68%           95%  
## 0.04733638 2.00227280
```

```
edgeR_results[1:5, 1:6]
```

```
##      gene LCGV.4.logFC LCGV.4.logCPM LCGV.4.LR LCGV.4.PValue  
## 1     A1BG  0.1939510    2.4093604 0.1224118    0.7264333  
## 2 A1BG-AS1 -0.3227594    3.4936307 0.9826982    0.3215336  
## 3     A1CF  0.0000000   -0.9245545 0.0000000    1.0000000  
## 4      A2M  0.0000000   -0.9245545 0.0000000    1.0000000  
## 5  A2M-AS1  0.6333559   -0.4253955 0.2134882    0.6440470  
## LCGV.4.PValue.adj  
## 1           1.0000000  
## 2           0.8634145  
## 3           1.0000000  
## 4           1.0000000  
## 5           1.0000000
```

```
row.names(edgeR_results) <- edgeR_results$gene  
dplyr::select(edgeR_results[edgeR_results$gene %in%
```

```

c("EGFR", "EP300", "HDAC4", "CDKN1A"), ], contains("M")) %>%
dplyr::select(contains("PValue.adj"))

##          LCM.4.PValue.adj LJM.4.PValue.adj LIM.4.PValue.adj GVM.4.PValue.adj
## CDKN1A      7.765397e-10   7.132150e-06   5.904157e-11   0.0004304325
## EGFR        1.000000e+00   6.757156e-01   5.802838e-01   0.4210450337
## EP300       9.065477e-01   1.000000e+00   5.297162e-01   1.0000000000
## HDAC4       1.000000e+00   1.000000e+00   7.308229e-01   1.0000000000
##          LCM.13.PValue.adj LJM.13.PValue.adj LIM.13.PValue.adj
## CDKN1A      1.098515e-32   2.310890e-15   1
## EGFR        1.073840e-09   2.314689e-05   1
## EP300       1.352439e-05   6.697016e-02   1
## HDAC4       1.539614e-06   9.621443e-02   1
##          GVM.13.PValue.adj LCM.22.PValue.adj LJM.22.PValue.adj
## CDKN1A      0.7364812     2.587257e-21   1.154165e-02
## EGFR        1.0000000     8.637272e-35   5.610384e-07
## EP300       1.0000000     2.969729e-05   8.009564e-01
## HDAC4       1.0000000     8.871445e-05   1.127586e-01
##          LIM.22.PValue.adj GVM.22.PValue.adj
## CDKN1A      0.5303285     0.8958338
## EGFR        0.5905420     1.0000000
## EP300       0.7554520     1.0000000
## HDAC4       0.5601210     1.0000000

write.csv(edgeR_results, paste0(thesis_tables_directory, TABLE_EDGER_RESULTS),
quote = F, row.names = F)
#write.table(edgeR_results, paste0(thesis_tables_directory, "TABLE_A11.txt"),
quote = F, row.names = F, sep = "\t")
#write.csv(dplyr::select(edgeR_results, c(gene, ends_with("LogFC"), ends_with("P
Value"))), paste0(root_directory, "edgeR_results_LCPval.csv"), quote =
F, row.names = F)

## Clean up the table containg the DE expresison information
edgeR_results.melt<-melt(edgeR_results)

## Using gene as id variables

edgeR_results.melt$variable<-gsub(pattern = "PValue.adj", replacement =
"Pvalue_adj", x = edgeR_results.melt$variable)
edgeR_results.melt<-separate(edgeR_results.melt, "variable", sep = "\\.", into =
c("comparison", "ExposureTime", "value_type"))

## Assign colors to BCS
color_map<-
c("LCGV"=unname(subject_long_taxa_colors["Lactobacillus_crispatus"]), "LCM"=un
name(subject_long_taxa_colors["Lactobacillus_crispatus"]), "LJGV"=unname(subje
ct_long_taxa_colors["Lactobacillus_jensenii"]), "LJM"=unname(subject_long_taxa
_colors["Lactobacillus_jensenii"]), "LIM"=unname(subject_long_taxa_colors["Lac
tobacillus_iners"]), "LIGV"=unname(subject_long_taxa_colors["Lactobacillus_ine
rs"]), "GVM"=unname(subject_long_taxa_colors["Gardnerella_vaginalis"]))

```

```

## Use different line types for G. vaginalis and medium (if used)
line_map<-
data.frame(comparison=names(color_map),line_type=c("solid","dashed","solid",
dashed","dashed","solid","dashed"))

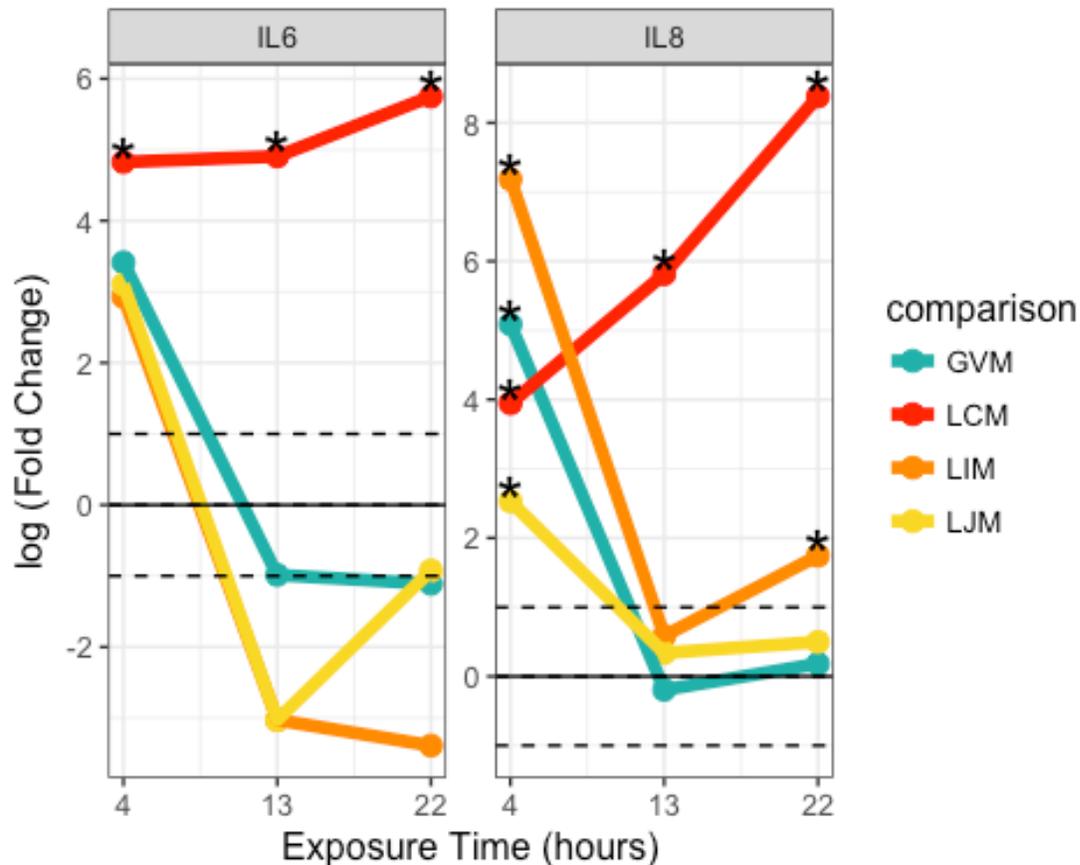
## Merge tables together to map DE results to plot annotations
edgeR_results_colors<-
data.frame(comparison=names(color_map),colr=unname(color_map))
edgeR_results.melt<-merge(edgeR_results.melt,edgeR_results_colors)
edgeR_results.melt<-merge(edgeR_results.melt,line_map)
edgeR_results.melt<-
merge(edgeR_results.melt,dplyr::select(Pathways_design,c(comp,L,ref)),all.x=T
,by.x="comparison",by.y="comp")
edgeR_results.melt$ExposureTime<-as.numeric(edgeR_results.melt$ExposureTime)

## Format and plot the selected immune gene expression's logFC over the
timecourse
immune_genes<-c("IL6","CXCL8")
immune_genes_expression<- dplyr::filter(edgeR_results.melt,gene %in%
immune_genes & ref=="M" ) ## Only include selected immune related genes vs.
the cell culture medium reference
immune_genes_expression<-unique(immune_genes_expression)
immune_genes_expression<-spread(immune_genes_expression,key =
value_type,value = value) ## This will make the logFC, FDR, and other DE
attributes into colums for easier plotting. The plot will use logFC and FDR
information
immune_genes_expression[immune_genes_expression$Pvalue_adj<0.01,"DE.pval"]<-
"*" ## Annotate which samples are DE by FDR
immune_genes_expression$gene<-factor(immune_genes_expression$gene,levels =
c("IL6","CXCL8"),ordered = T,labels = c("IL6","IL8")) ## Make plot consistent

##Make plot
long_plot.immune<-
ggplot(immune_genes_expression)+geom_point(aes(x=as.numeric(ExposureTime),y=logFC,col=comparison),size=3)+
geom_line(aes(x=as.numeric(ExposureTime),y=logFC,col=comparison),size=2)+
facet_wrap(~gene,scales = "free_y",ncol
=2)+theme_bw()+scale_color_manual(values = color_map)+
geom_hline(yintercept =0)+geom_hline(yintercept =c(-
1,0,1),lty=2)+xlab("Exposure Time (hours)")+
geom_text(aes(x=as.numeric(ExposureTime),y=logFC,label=DE.pval),size=8)+
mBio+
ylab("log (Fold Change)")+scale_x_continuous(breaks=c(4,13,22))

plot(long_plot.immune)

```



```

cairo_ps(paste0(thesis_figures_directory,FIGURE_LONGITUDINAL_GENEEXP.immune),
width = 8,height = 6)
plot(long_plot.immune)
dev.off()

## quartz_off_screen
##          2

## Format and plot the cell cycle pathway-related gene expression's logFC
over the timecourse
cycle_genes<-c("HDAC4","EP300","CDKN1A","CDK4","CCND1","CCNE2","ESR1","EGFR")
## Select which cell cycle genes to plot
cycle_genes_expression<- dplyr::filter(edgeR_results.melt,gene %in%
cycle_genes & ref=="M" ) ## Only include the cell cycle genes and cell
culture medium as the reference
cycle_genes_expression<-unique(cycle_genes_expression)
cycle_genes_expression<-spread(cycle_genes_expression,key = value_type,value
= value) ## This will make the logFC, FDR, and other DE attributes into
columns for easier plotting. The plot will use logFC and FDR information
cycle_genes_expression$gene<-factor(cycle_genes_expression$gene,levels =
cycle_genes,ordered = T) ## Maintain order of genes- this follows logical
order discussed in thesis.
cycle_genes_expression[cycle_genes_expression$Pvalue_adj<0.01,"DE.pval"]<- "*"

```

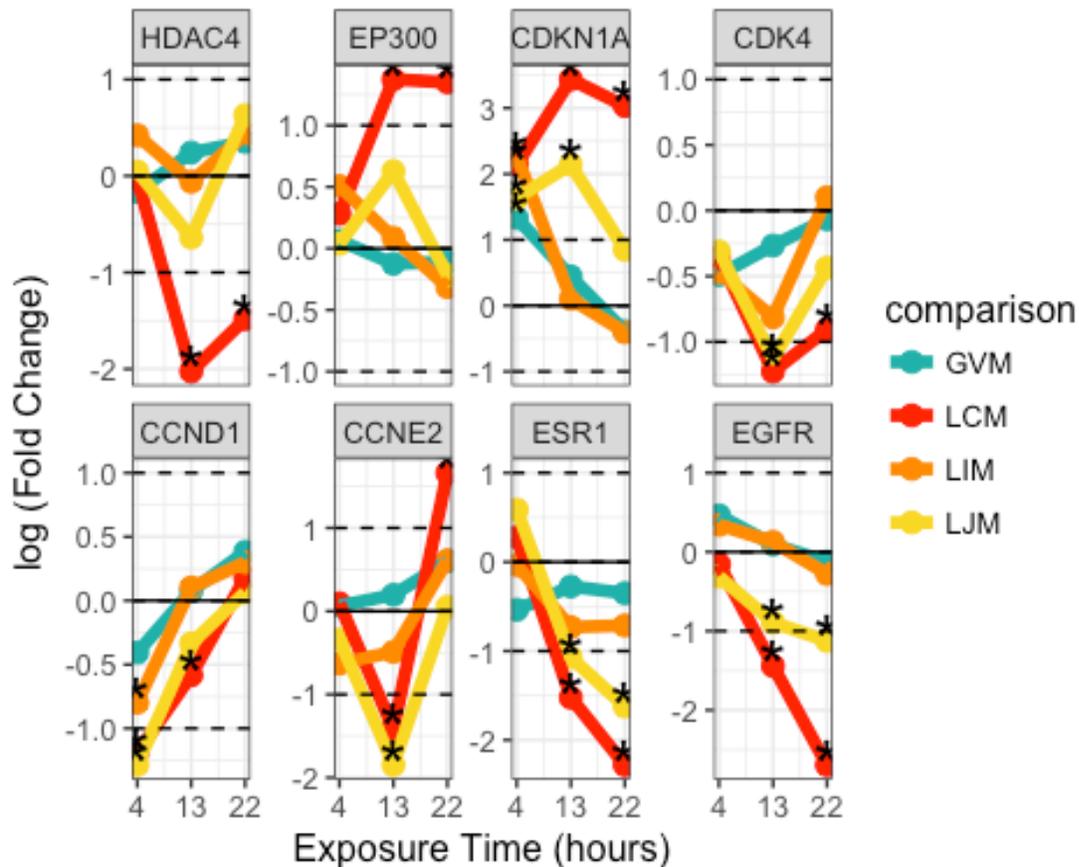
```

## Annotate which samples are DE by FDR

## Make plot
long_plot.cc<-
ggplot(cycle_genes_expression)+geom_point(aes(x=as.numeric(ExposureTime),y=log
FC,col=comparison),size=3)+
  geom_line(aes(x=as.numeric(ExposureTime),y=logFC,col=comparison),size=2)+
  facet_wrap(~gene,scales = "free_y",nrow
=2)+theme_bw()+scale_color_manual(values = color_map)+
  geom_hline(yintercept =0)+geom_hline(yintercept =c(-
1,0,1),lty=2)+xlab("Exposure Time (hours)")+
  geom_text(aes(x=as.numeric(ExposureTime),y=logFC,label=DE.pval),size=8)+
  mBio+
  ylab("log (Fold Change)")+scale_x_continuous(breaks=c(4,13,22))

plot(long_plot.cc)

```



```

cairo_ps(paste0(thesis_figures_directory,FIGURE_LONGITUDINAL_GENEEXP.cycle),w
idth = 8,height = 6)
plot(long_plot.cc)
dev.off()

```

End Timestamp

```
## Log session info
filewritable_time<-gsub(gsub(Sys.time(),pattern = " ",replacement =
"_"),pattern = "-|:",replacement = "")
(sessionInfo_latex<-toLatex(sessionInfo()))

## \begin{itemize}\raggedright
## \item R version 3.3.1 (2016-06-21), \verb|x86_64-apple-darwin13.4.0|
## \item Locale: \verb|en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-
8/en_US.UTF-8|
## \item Base packages: base, datasets, graphics, grDevices, grid,
## methods, parallel, stats, utils
## \item Other packages: Biobase~2.34.0, BiocGenerics~0.20.0,
## Boruta~5.2.0, caret~6.0-76, dplyr~0.5.0, edgeR~3.16.5,
## ggbiplot~0.55, ggplot2~2.2.1, gPCA~1.0, gplots~3.0.1,
## gridExtra~2.2.1, lattice~0.20-35, limma~3.30.13, nlme~3.1-131,
## plotly~4.6.0, plyr~1.8.4, psych~1.7.3.21, purrr~0.2.2,
## randomForest~4.6-12, ranger~0.7.0, RColorBrewer~1.1-2,
## readr~1.1.0, reshape~0.8.6, rfPermute~2.1.5, scales~0.4.1,
## squash~1.0.7, stringr~1.2.0, tibble~1.3.0, tidyr~0.6.1,
## tidyverse~1.1.1
## \item Loaded via a namespace (and not attached): abind~1.4-5,
## assertthat~0.2.0, backports~1.0.5, bitops~1.0-6, broom~0.4.2,
## car~2.1-4, caTools~1.17.1, cellranger~1.1.0, codetools~0.2-15,
## colorspace~1.3-2, DBI~0.6-1, deldir~0.1-14, digest~0.6.12,
## evaluate~0.10, forcats~0.2.0, foreach~1.4.3, foreign~0.8-68,
## gdata~2.17.0, goftest~1.1-1, gtable~0.2.0, gtools~3.5.0,
## haven~1.0.0, hms~0.3, htmltools~0.3.5, htmlwidgets~0.8,
## httr~1.2.1, iterators~1.0.8, jsonlite~1.4, KernSmooth~2.23-15,
## knitr~1.15.1, labeling~0.3, lazyeval~0.2.0, lme4~1.1-13,
## locfit~1.5-9.1, lubridate~1.6.0, magrittr~1.5, mapdata~2.2-6,
## maps~3.1.1, MASS~7.3-47, Matrix~1.2-8, MatrixModels~0.4-1,
## mgcv~1.8-17, minqa~1.2.4, mnormt~1.5-5, ModelMetrics~1.1.0,
## modelr~0.1.0, munsell~0.4.3, nloptr~1.0.4, nnet~7.3-12,
## pbkrtest~0.4-7, polyclip~1.6-1, quantreg~5.33, R6~2.2.0,
## Rcpp~0.12.10, readxl~1.0.0, reshape2~1.4.2, rmarkdown~1.5,
## rpart~4.1-11, rprojroot~1.2, rvest~0.3.2, SparseM~1.77,
## spatstat~1.50-0, spatstat.utils~1.4-1, splines~3.3.1,
## stats4~3.3.1, stringi~1.1.5, swfscMisc~1.2, tensor~1.5,
## tools~3.3.1, viridisLite~0.2.0, xml2~1.1.1, yaml~2.1.14
## \end{itemize}

write(sessionInfo_latex,paste0(R_script_output_directory,"R_sessionInfo_",fil
ewritable_time, ".log.txt"))

timestamp()

## ##----- Wed May 31 23:54:34 2017 -----##
```

Appendix 3 Cell Line Authentication Forms

Report Date: 11/07/2016

Sample (s) received Date: 11/04/2016

Name of Requester: Steven Smith

Cell Line ID: A2EN PSS

Biopolymer Lab Order: GCF-SS-1050

CELL LINE AUTHENTICATION REPORT

RESEARCH USE ONLY

Allele Table

Locus	A2EN PSS
Amelogenin	X
CSF1PO	12,13
D13S317	9,13
D16S539	9,11
D21S11	29,32.2
D5S818	11,13
D7S820	11
TH01	6
TPOX	8,11
vWA	17

Interpretation

Sample SS-A2EN PSS did not match with any ATCC Reference.

Explanation of Test Results

Cell lines with $\geq 80\%$ match are considered to be related; i.e., derived from a common ancestry. A cell line with an STR profile match of $\leq 56\%$ is considered unrelated. A unique cell line has a STR profile that is different from another unique cell line.

- The submitted sample profile is human, but not a match for any profile in the ATCC STR database.
- The submitted profile is an exact match for the following ATCC human cell line(s) in the ATCC STR database (8 core loci plus Amelogenin):
- The submitted profile is similar to the following ATCC human cell line(s).
- The submitted cell line is contaminated, and has tested positive for mouse marker.

STR typing was performed using the Promega GenePrint 10 System™. The kit includes 10 human specific loci for human cell line authentication. The human loci collectively provide a genetic profile with a random match probability of 1 in 2.92×10^9 . Where more than one human profile is observed, alleles of the minor contributor are indicated in parentheses. . Our laboratory uses GenePrint® 5X Mouse Primer Pair Mix is designed to be used as a sensitive marker that specifically detects the presence of mouse (*Mus musculus*) DNA while simultaneously providing detection of about 1% fraction of mouse contaminant in a human cell line when using extracted DNA.

Report Date: 11/07/2016

Sample (s) received Date: 11/04/2016

Name of Requester: Steven Smith

Cell Line ID: A2EN P62

Biopolymer Lab Order: GCF-SS-1050

CELL LINE AUTHENTICATION REPORT

RESEARCH USE ONLY

Allele Table

Locus	A2EN P62
Amelogenin	X
CSF1PO	12,13
D13S317	9,13
D16S539	9,11
D21S11	29,32.2
D5S818	11,13
D7S820	11
TH01	6
TPOX	8,11
vWA	17

Interpretation

Sample SS-A2EN P62 did not match with any ATCC Reference.

Explanation of Test Results

Cell lines with $\geq 80\%$ match are considered to be related; i.e., derived from a common ancestry. A cell line with an STR profile match of $\leq 56\%$ is considered unrelated. A unique cell line has a STR profile that is different from another unique cell line.

- The submitted sample profile is human, but not a match for any profile in the ATCC STR database.
- The submitted profile is an exact match for the following ATCC human cell line(s) in the ATCC STR database (8 core loci plus Amelogenin):
- The submitted profile is similar to the following ATCC human cell line(s).
- The submitted cell line is contaminated, and has tested positive for mouse marker.

STR typing was performed using the Promega GenePrint 10 System™. The kit includes 10 human specific loci for human cell line authentication. The human loci collectively provide a genetic profile with a random match probability of 1 in 2.92×10^9 . Where more than one human profile is observed, alleles of the minor contributor are indicated in parentheses. . Our laboratory uses GenePrint® 5X Mouse Primer Pair Mix is designed to be used as a sensitive marker that specifically detects the presence of mouse (*Mus musculus*) DNA while simultaneously providing detection of about 1% fraction of mouse contaminant in a human cell line when using extracted DNA.
Report Date: 11/07/2016

Sample (s) received Date: 11/04/2016

Name of Requester: Steven Smith

Cell Line ID: VK2 P31

Biopolymer Lab Order: GCF-SS-1050

CELL LINE AUTHENTICATION REPORT

RESEARCH USE ONLY

Allele Table

Locus	VK2 P31	ATCC Reference CRL-2616(VK2)
Amelogenin	X	X
CSF1PO	10,11	10,11
D13S317	9,12	9,12
D16S539	9	9
D21S11	29,31.2	NO DATA
D5S818	9,10	9,10
D7S820	10,11	10,11
TH01	7,9.3	7,9.3
TPOX	11	11
vWA	16	16

Interpretation

The sample provided exhibits identical genetic profiles.

Sample VK2 P31 shares 13 alleles of 13 alleles (100.0 %) with the ATCC reference.

Explanation of Test Results

Cell lines with $\geq 80\%$ match are considered to be related; i.e., derived from a common ancestry. A cell line with an STR profile match of $\leq 56\%$ is considered unrelated. A unique cell line has a STR profile that is different from another unique cell line.

- The submitted sample profile is human, but not a match for any profile in the ATCC STR database.
- The submitted profile is an exact match for the following ATCC human cell line(s) in the ATCC STR database (8 core loci plus Amelogenin): **CRL-2616(VK2)**
- The submitted profile is similar to the following ATCC human cell line(s).
- The submitted cell line is contaminated, and has tested positive for mouse marker.

STR typing was performed using the Promega GenePrint 10 System™. The kit includes 10 human specific loci for human cell line authentication. The human loci collectively provide a genetic profile with a random match probability of 1 in 2.92×10^9 . Where more than one human profile is observed, alleles of the minor contributor are indicated in parentheses. . Our laboratory uses GenePrint® 5X Mouse Primer Pair Mix is designed to be used as a sensitive marker that specifically detects the presence of mouse (*Mus musculus*) DNA while simultaneously providing detection of about 1% fraction of mouse contaminant in a human cell line when using extracted DNA.

Appendix 4 Small RNA-seq miRNA raw read counts table

Small RNA-seq miRNA raw read counts table for each sequenced sample from Figure 2.1 available at https://github.com/ravel-lab/smith_thesis_2017/tree/master/AnalysisPipeline/Tables/ as TABLE_A4.csv

Appendix 5 Post-QC log2 normalized small RNA-seq counts and metadata

Post-QC log2 normalized small RNA-seq counts and metadata used to test and train proxy-Amsel-RF and Nugent-RF of samples used in Figure 2.1 and Figure 2.3 available at https://github.com/ravel-lab/smith_thesis_2017/tree/master/AnalysisPipeline/Tables/ as TABLE_A5.csv

Appendix 6 Small RNA-seq library & subject metadata

Small RNA-seq library preparation, sequencing & alignment statistics, QC analysis annotation & subject metadata used in analysis of samples in Figure 2.1 and corresponding data dictionary available at https://github.com/ravel-lab/smith_thesis_2017/tree/master/AnalysisPipeline/Tables/ as TABLE_A6.csv

Appendix 7 Metataxonomic data and metadata used to train and test the Amsel

Random Forest model

Metataxonomic data and metadata used to train and test the Amsel Random Forest model available at https://github.com/ravel-lab/smith_thesis_2017/tree/master/AnalysisPipeline/Tables/ as TABLE_A7.csv

Appendix 8 Importance metrics and p-values for the Amsel Random Forest variable selection results

Importance metrics and p-values for the Amsel Random Forest variable selection results available at https://github.com/ravel-lab/smith_thesis_2017/tree/master/AnalysisPipeline/Tables/ as TABLE_A8.csv

Appendix 9 Metataxonomic data and metadata used as inputs to classify samples for proxy-Amsel diagnosis

Metataxonomic data and metadata used as inputs to classify samples for proxy-Amsel diagnosis available at https://github.com/ravel-lab/smith_thesis_2017/tree/master/AnalysisPipeline/Tables/ as TABLE_A9.csv

Appendix 10 Importance metrics and p-values for the proxy-Amsel-RF and Nugent-RF variable selection results

Importance metrics and p-values for the proxy-Amsel-RF and Nugent-RF variable selection results from Figure 2.3 available at https://github.com/ravel-lab/smith_thesis_2017/tree/master/AnalysisPipeline/Tables/ as TABLE_A10.csv

Appendix 11 rRNA-reduced RNA-seq raw read counts table

rRNA-reduced RNA-seq raw read counts table available at https://github.com/ravel-lab/smith_thesis_2017/tree/master/AnalysisPipeline/Tables/ as TABLE_A11.csv

Appendix 12 edgeR GLM-based LRT results

edgeR GLM-based LRT results for differential expression analysis and IPA analysis

available at <https://github.com/ravel->

[lab/smith_thesis_2017/tree/master/AnalysisPipeline/Tables/](https://github.com/ravel-lab/smith_thesis_2017/tree/master/AnalysisPipeline/Tables/) as TABLE_A12.csv

Appendix 13 Canonical pathway z-scores

IPA-based activation score (z-scores) for each exposure time BCS vs. cell culture

medium available at <https://github.com/ravel->

[lab/smith_thesis_2017/tree/master/AnalysisPipeline/Tables/](https://github.com/ravel-lab/smith_thesis_2017/tree/master/AnalysisPipeline/Tables/) as TABLE_A13.csv

Bibliography

- [1] A. L. Kau, P. P. Ahern, N. W. Griffin, A. L. Goodman, and J. I. Gordon, "Human nutrition, the gut microbiome and the immune system," *Nature*, vol. 474, no. 7351, pp. 327–336, Jun. 2011.
- [2] J. Chen, Y. Li, Y. Tian, C. Huang, D. Li, Q. Zhong, and X. Ma, "Interaction between Microbes and Host Intestinal Health: Modulation by Dietary Nutrients and Gut-Brain-Endocrine-Immune Axis.," *Curr. Protein Pept. Sci.*, vol. 16, no. 7, pp. 592–603, 2015.
- [3] M. Conlon and A. Bird, "The Impact of Diet and Lifestyle on Gut Microbiota and Human Health," *Nutrients*, vol. 7, no. 1, pp. 17–44, Jan. 2015.
- [4] M. E. Wiens and J. G. Smith, "Alpha-defensin HD5 inhibits furin cleavage of human papillomavirus 16 L2 to block infection.," *Journal of Virology*, vol. 89, no. 5, pp. 2866–2874, Mar. 2015.
- [5] M. Farage and H. Maibach, "Lifetime changes in the vulva and vagina," *Arch Gynecol Obstet*, vol. 273, no. 4, pp. 195–202, Oct. 2005.
- [6] J. Ravel, P. Gajer, Z. Abdo, G. M. Schneider, S. S. K. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, R. M. Brotman, C. C. Davis, K. Ault, L. Peralta, and L. J. Forney, "Vaginal microbiome of reproductive-age women.," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 108, pp. 4680–4687, Mar. 2011.
- [7] P. Gajer, R. M. Brotman, G. Bai, J. Sakamoto, U. M. E. Schütte, X. Zhong, S. S. K. Koenig, L. Fu, Z. S. Ma, X. Zhou, Z. Abdo, L. J. Forney, and J. Ravel, "Temporal dynamics of the human vaginal microbiota.," *Sci Transl Med*, vol. 4, no. 132, p. 132ra52, May 2012.
- [8] X. Zhou, C. J. Brown, Z. Abdo, C. C. Davis, M. A. Hansmann, P. Joyce, J. A. Foster, and L. J. Forney, "Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women.," *ISME J*, vol. 1, no. 2, pp. 121–133, Jun. 2007.
- [9] D. N. Fredricks, T. L. Fiedler, and J. M. Marrazzo, "Molecular identification of bacteria associated with bacterial vaginosis.," *N. Engl. J. Med.*, vol. 353, no. 18, pp. 1899–1911, Nov. 2005.
- [10] A. C. C. Campos, R. Freitas-Junior, L. F. J. Ribeiro, R. R. Paulinelli, and C. Reis, "Prevalence of vulvovaginitis and bacterial vaginosis in patients with koilocytosis.," *Sao Paulo Med J*, vol. 126, no. 6, pp. 333–336, Nov. 2008.
- [11] R. P. Nugent, M. A. Krohn, and S. L. Hillier, "Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation.," *J. Clin. Microbiol.*, vol. 29, no. 2, pp. 297–301, Feb. 1991.
- [12] D. H. Martin, "The microbiota of the vagina and its influence on women's health and disease," *The American journal of the medical sciences*, vol. 343, no. 1, pp. 2–9, 2012.
- [13] R. Amsel, P. A. Totten, C. A. Spiegel, K. C. Chen, D. Eschenbach, and K. K. Holmes, "Nonspecific vaginitis. Diagnostic criteria and microbial and epidemiologic associations.," *Am. J. Med.*, vol. 74, no. 1, pp. 14–22, Jan. 1983.
- [14] P. E. Hay, A. Ugwumadu, and J. Chowns, "Sex, thrush and bacterial

- vaginosis.,” *International Journal of STD & AIDS*, vol. 8, no. 10, pp. 603–608, Oct. 1997.
- [15] J. R. Schwebke, C. M. Richey, and H. L. Weiss², “Correlation of behaviors with microbiological changes in vaginal flora.,” *J INFECT DIS*, vol. 180, no. 5, pp. 1632–1636, Nov. 1999.
- [16] F. E. Keane, C. A. Ison, and D. Taylor-Robinson, “A longitudinal study of the vaginal flora over a menstrual cycle.,” *International Journal of STD & AIDS*, vol. 8, no. 8, pp. 489–494, Aug. 1997.
- [17] R. M. Brotman, J. Ravel, R. A. Cone, and J. M. Zenilman, “Rapid fluctuation of the vaginal microbiota measured by Gram stain analysis.,” *Sex Transm Infect*, vol. 86, no. 4, pp. 297–302, Aug. 2010.
- [18] R. M. Brotman, K. G. Ghanem, M. A. Klebanoff, T. E. Taha, D. O. Scharfstein, and J. M. Zenilman, “The effect of vaginal douching cessation on bacterial vaginosis: a pilot study.,” *American Journal of Obstetrics and Gynecology*, vol. 198, no. 6, pp. 628.e1–7, Jun. 2008.
- [19] J. Ravel, R. M. Brotman, P. Gajer, B. Ma, M. Nandy, D. W. Fadrosh, J. Sakamoto, S. S. Koenig, L. Fu, X. Zhou, R. J. Hickey, J. R. Schwebke, and L. J. Forney, “Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis,” *Microbiome*, vol. 1, no. 1, p. 29, 2013.
- [20] S. Srinivasan, C. Liu, C. M. Mitchell, T. L. Fiedler, K. K. Thomas, K. J. Agnew, J. M. Marrazzo, and D. N. Fredricks, “Temporal variability of human vaginal bacteria and relationship with bacterial vaginosis.,” *PLoS ONE*, vol. 5, no. 4, p. e10197, 2010.
- [21] I. M. Linhares, P. R. Summers, B. Larsen, P. C. Giraldo, and S. S. Witkin, “Contemporary perspectives on vaginal pH and lactobacilli.,” *American Journal of Obstetrics and Gynecology*, vol. 204, no. 2, pp. 120.e1–5, Feb. 2011.
- [22] M. I. Petrova, E. Lievens, S. Malik, N. Imholz, and S. Lebeer, “Lactobacillus species as biomarkers and agents that can promote various aspects of vaginal health,” *Front. Physiol.*, vol. 6, Mar. 2015.
- [23] B. Ma, L. J. Forney, and J. Ravel, “Vaginal microbiome: rethinking health and disease.,” *Annu. Rev. Microbiol.*, vol. 66, pp. 371–389, 2012.
- [24] G. Stoyancheva, M. Marzotto, F. Dellaglio, and S. Torriani, “Bacteriocin production and gene sequencing analysis from vaginal Lactobacillus strains.,” *Arch. Microbiol.*, vol. 196, no. 9, pp. 645–653, Sep. 2014.
- [25] J. Zheng, M. G. Gänzle, X. B. Lin, L. Ruan, and M. Sun, “Diversity and dynamics of bacteriocins from human microbiome,” *Environ. Microbiol.*, vol. 17, no. 6, pp. 2133–2143, Jun. 2015.
- [26] M. N. Anahtar, E. H. Byrne, K. E. Doherty, B. A. Bowman, H. S. Yamamoto, M. Soumillon, N. Padavattan, N. Ismail, A. Moodley, M. E. Sabatini, M. S. Ghebremichael, C. Nusbaum, C. Huttenhower, H. W. Virgin, T. Ndung’u, K. L. Dong, B. D. Walker, R. N. Fichorova, and D. S. Kwon, “Cervicovaginal Bacteria Are a Major Modulator of Host Inflammatory Responses in the Female Genital Tract,” *Immunity*, vol. 42, no. 5, pp. 965–976, May 2015.

- [27] X. Zhou, S. J. Bent, M. G. Schneider, C. C. Davis, M. R. Islam, and L. J. Forney, "Characterization of vaginal microbial communities in adult healthy women using cultivation-independent methods.," *Microbiology (Reading, Engl.)*, vol. 150, no. 8, pp. 2565–2573, Aug. 2004.
- [28] T. L. Chernes, L. A. Meyn, M. A. Krohn, J. G. Lurie, and S. L. Hillier, "Association between acquisition of herpes simplex virus type 2 in women and bacterial vaginosis.," *Clin. Infect. Dis.*, vol. 37, no. 3, pp. 319–325, Aug. 2003.
- [29] H. L. Martin, B. A. Richardson, P. M. Nyange, L. Lavreys, S. L. Hillier, B. Chohan, K. Mandaliya, J. O. Ndinya-Achola, J. Bwayo, and J. Kreiss, "Vaginal lactobacilli, microbial flora, and risk of human immunodeficiency virus type 1 and sexually transmitted disease acquisition.," *J INFECT DIS*, vol. 180, no. 6, pp. 1863–1868, Dec. 1999.
- [30] S. E. Peters, C. M. Beck-Sagué, C. E. Farshy, I. Gibson, K. A. Kubota, F. Solomon, S. A. Morse, A. J. Sievert, and C. M. Black, "Behaviors associated with *Neisseria gonorrhoeae* and *Chlamydia trachomatis*: cervical infection among young women attending adolescent clinics.," *Clin Pediatr (Phila)*, vol. 39, no. 3, pp. 173–177, Mar. 2000.
- [31] C. R. Cohen, A. Duerr, N. Pruthithada, S. Ruggao, S. Hillier, P. Garcia, and K. Nelson, "Bacterial vaginosis and HIV seroprevalence among female commercial sex workers in Chiang Mai, Thailand.," *AIDS*, vol. 9, no. 9, pp. 1093–1097, Sep. 1995.
- [32] T. E. Taha, D. R. Hoover, G. A. Dallabetta, N. I. Kumwenda, L. A. Mtimavalye, L. P. Yang, G. N. Liomba, R. L. Broadhead, J. D. Chipangwi, and P. G. Miotti, "Bacterial vaginosis and disturbances of vaginal flora: association with increased acquisition of HIV.," *AIDS*, vol. 12, no. 13, pp. 1699–1706, Sep. 1998.
- [33] S. Cu-Uvin, J. W. Hogan, A. M. Caliendo, J. Harwell, K. H. Mayer, C. C. Carpenter, HIV Epidemiology Research Study, "Association between bacterial vaginosis and expression of human immunodeficiency virus type 1 RNA in the female genital tract.," *Clin. Infect. Dis.*, vol. 33, no. 6, pp. 894–896, Sep. 2001.
- [34] J. S. Coleman, J. Hitti, E. A. Bukusi, C. Mwachari, A. Muliro, R. Nguti, R. Gausman, S. Jensen, D. Patton, D. Lockhart, R. Coombs, and C. R. Cohen, "Infectious correlates of HIV-1 shedding in the female upper and lower genital tracts.," *AIDS*, vol. 21, no. 6, pp. 755–759, Mar. 2007.
- [35] R. L. Goldenberg, W. W. Andrews, A. C. Yuan, H. T. MacKay, and M. E. St Louis, "Sexually transmitted diseases and adverse outcomes of pregnancy.," *Clin Perinatol*, vol. 24, no. 1, pp. 23–41, Mar. 1997.
- [36] M. G. Gravett, H. P. Nelson, T. DeRouen, C. Critchlow, D. A. Eschenbach, and K. K. Holmes, "Independent associations of bacterial vaginosis and *Chlamydia trachomatis* infection with adverse pregnancy outcome.," *JAMA*, vol. 256, no. 14, pp. 1899–1903, Oct. 1986.
- [37] S. L. Hillier, R. P. Nugent, D. A. Eschenbach, M. A. Krohn, R. S. Gibbs, D. H. Martin, M. F. Cotch, R. Edelman, J. G. Pastorek, and A. V. Rao, "Association between bacterial vaginosis and preterm delivery of a low-

- birth-weight infant. The Vaginal Infections and Prematurity Study Group.," *N. Engl. J. Med.*, vol. 333, no. 26, pp. 1737–1742, Dec. 1995.
- [38] H. M. McDonald, J. A. O'Loughlin, P. Jolley, R. Vigneswaran, and P. J. McDonald, "Prenatal microbiological risk factors associated with preterm birth.," *Br J Obstet Gynaecol*, vol. 99, no. 3, pp. 190–196, Mar. 1992.
- [39] P. J. Meis, R. L. Goldenberg, B. Mercer, A. Moawad, A. Das, D. McNellis, F. Johnson, J. D. Iams, E. Thom, and W. W. Andrews, "The preterm prediction study: significance of vaginal infections. National Institute of Child Health and Human Development Maternal-Fetal Medicine Units Network.," *YMOB*, vol. 173, no. 4, pp. 1231–1235, Oct. 1995.
- [40] A. Aiyar, A. J. Quayle, L. R. Buckner, S. P. Sherchand, T. L. Chang, A. H. Zea, D. H. Martin, and R. J. Belland, "Influence of the tryptophan-indole-IFN γ axis on human genital Chlamydia trachomatis infection: role of vaginal co-infections.," *Front Cell Infect Microbiol*, vol. 4, p. 72, 2014.
- [41] R. B. Pyles, K. L. Vincent, M. M. Baum, B. Elsom, A. L. Miller, C. Maxwell, T. D. Eaves-Pyles, G. Li, V. L. Popov, R. J. Nusbaum, and M. R. Ferguson, "Cultivated Vaginal Microbiomes Alter HIV-1 Infection and Antiretroviral Efficacy in Colonized Epithelial Multilayer Cultures.," *PLoS ONE*, vol. 9, no. 3, p. e93419, 2014.
- [42] J. M. Macklaim, A. D. Fernandes, J. M. Di Bella, J.-A. Hammond, G. Reid, and G. B. Gloor, "Comparative meta-RNA-seq of the vaginal microbiota and differential expression by Lactobacillus iners in health and dysbiosis.," *Microbiome*, vol. 1, no. 1, p. 12, 2013.
- [43] J. M. Macklaim, G. B. Gloor, K. C. Anukam, S. Cribby, and G. Reid, "At the crossroads of vaginal health and disease, the genome sequence of Lactobacillus iners AB-1.," *Proceedings of the National Academy of Sciences*, vol. 108, pp. 4688–4695, Mar. 2011.
- [44] C. R. Wira, J. V. Fahey, C. L. Sentman, P. A. Pioli, and L. Shen, "Innate and adaptive immunity in female genital tract: cellular responses and interactions.," *Immunol. Rev.*, vol. 206, no. 1, pp. 306–335, Aug. 2005.
- [45] B. Usluogullari, I. Gumus, E. Gunduz, I. Kaygusuz, S. Simavli, M. Acar, M. Oznur, M. Gunduz, and H. Kafali, "The role of Human Dectin-1 Y238X Gene Polymorphism in recurrent vulvovaginal candidiasis infections.," *Mol. Biol. Rep.*, vol. 41, no. 10, pp. 6763–6768, Oct. 2014.
- [46] A. Carvalho, G. Giovannini, A. De Luca, C. D'Angelo, A. Casagrande, R. G. Iannitti, G. Ricci, C. Cunha, and L. Romani, "Dectin-1 isoforms contribute to distinct Th1/Th17 cell activation in mucosal candidiasis.," *Cell. Mol. Immunol.*, vol. 9, no. 3, pp. 276–286, May 2012.
- [47] C. Mitchell and J. Marrazzo, "Bacterial Vaginosis and the Cervicovaginal Immune Response," *American Journal of Reproductive Immunology*, vol. 71, no. 6, pp. 555–563, May 2014.
- [48] S. S. Witkin, I. M. Linhares, and P. Giraldo, "Bacterial flora of the female genital tract: function and immune regulation.," *Best Pract Res Clin Obstet Gynaecol*, vol. 21, no. 3, pp. 347–354, Jun. 2007.
- [49] A. W. Horne, S. J. Stock, and A. E. King, "Innate immunity and disorders of the female reproductive tract.," *Reproduction*, vol. 135, no. 6, pp. 739–

- 749, Jun. 2008.
- [50] R. M. Brotman, J. Ravel, P. M. Bavoil, P. E. Gravitt, and K. G. Ghanem, "Microbiome, sex hormones, and immune responses in the reproductive tract: Challenges for vaccine development against sexually transmitted infections," *Vaccine*, Oct. 2013.
- [51] P. V. Nguyen, J. K. Kafka, V. H. Ferreira, K. Roth, and C. Kaushic, "Innate and adaptive immune responses in male and female reproductive tracts in homeostasis and following HIV infection," *Cell. Mol. Immunol.*, vol. 11, no. 5, pp. 410–427, 2014.
- [52] M. R. Genç, S. Vardhana, M. L. Delaney, S. S. Witkin, A. B. Onderdonk, MAP Study Group, "TNFA-308G>A polymorphism influences the TNF-alpha response to altered vaginal flora.," *European Journal of Obstetrics and Gynecology*, vol. 134, no. 2, pp. 188–191, Oct. 2007.
- [53] P. C. Giraldo, O. Babula, A. K. S. Gonçalves, I. M. Linhares, R. L. Amaral, W. J. Ledger, and S. S. Witkin, "Mannose-binding lectin gene polymorphism, vulvovaginal candidiasis, and bacterial vaginosis.," *Obstet Gynecol*, vol. 109, no. 5, pp. 1123–1128, May 2007.
- [54] M. R. Genç, A. B. Onderdonk, S. Vardhana, M. L. Delaney, E. R. Norwitz, R. E. Tuomala, L.-R. Paraskevas, S. S. Witkin, MAP Study Group, "Polymorphism in intron 2 of the interleukin-1 receptor antagonist gene, local midtrimester cytokine response to vaginal flora, and subsequent preterm birth.," *American Journal of Obstetrics and Gynecology*, vol. 191, no. 4, pp. 1324–1330, Oct. 2004.
- [55] M. R. Genç, S. Vardhana, M. L. Delaney, A. Onderdonk, R. Tuomala, E. Norwitz, S. S. Witkin, MAP Study Group, "Relationship between a toll-like receptor-4 gene polymorphism, bacterial vaginosis-related flora and vaginal cytokine responses in pregnant women.," *European Journal of Obstetrics and Gynecology*, vol. 116, no. 2, pp. 152–156, Oct. 2004.
- [56] K. K. Ryckman, S. M. Williams, M. A. Krohn, and H. N. Simhan, "Interaction between interleukin-1 receptor 2 and Toll-like receptor 4, and cervical cytokines," *Journal of Reproductive Immunology*, vol. 90, no. 2, pp. 220–226, Aug. 2011.
- [57] R. D. Mackelprang, C. W. Scoville, C. R. Cohen, R. O. Ondondo, A. W. Bigham, C. Celum, M. S. Campbell, M. Essex, A. Wald, J. Kiarie, A. Ronald, G. Gray, and J. R. Lingappa, "Toll-like receptor gene variants and bacterial vaginosis among HIV-1 infected and uninfected African women," *Genes Immun*, Apr. 2015.
- [58] K. E. Royse, M.-C. Kempf, G. McGwin, C. M. Wilson, J. Tang, and S. Shrestha, "Toll-like receptor gene variants associated with bacterial vaginosis among HIV-1 infected adolescents.," *Journal of Reproductive Immunology*, vol. 96, no. 1, pp. 84–89, Dec. 2012.
- [59] W. A. Rose, C. L. McGowin, R. A. Spagnuolo, T. D. Eaves-Pyles, V. L. Popov, and R. B. Pyles, "Commensal bacteria modulate innate immune responses of vaginal epithelial cell multilayer cultures.," *PLoS ONE*, vol. 7, no. 3, p. e32728, 2012.
- [60] O. Neth, D. L. Jack, A. W. Dodds, H. Holzel, N. J. Klein, and M. W.

- Turner, “Mannose-binding lectin binds to a range of clinically relevant microorganisms and promotes complement deposition.,” *Infection and Immunity*, vol. 68, no. 2, pp. 688–693, Feb. 2000.
- [61] M. W. Turner, “The role of mannose-binding lectin in health and disease,” *Molecular Immunology*, vol. 40, no. 7, pp. 423–429, Nov. 2003.
- [62] Y.-Y. Wang, A. Kannan, K. L. Nunn, M. A. Murphy, D. B. Subramani, T. Moench, R. Cone, and S. K. Lai, “IgG in cervicovaginal mucus traps HSV and prevents vaginal Herpes infections,” *Mucosal Immunol*, vol. 7, no. 5, pp. 1036–1044, Feb. 2014.
- [63] E. C. Tramont, “Inhibition of adherence of *Neisseria gonorrhoeae* by human genital secretions.,” *J. Clin. Invest.*, vol. 59, no. 1, pp. 117–124, Jan. 1977.
- [64] V. L. Yarbrough, S. Winkle, and M. M. Herbst-Kralovetz, “Antimicrobial peptides in the female reproductive tract: a critical component of the mucosal immune barrier with physiological and clinical implications,” *Human Reproduction Update*, vol. 21, no. 3, pp. 353–377, Apr. 2015.
- [65] J. Ding, Y.-Y. Chou, and T. L. Chang, “Defensins in viral infections.,” *J Innate Immun*, vol. 1, no. 5, pp. 413–420, 2009.
- [66] S. S. Wilson, M. E. Wiens, and J. G. Smith, “Antiviral Mechanisms of Human Defensins,” *J. Mol. Biol.*, vol. 425, no. 24, pp. 4965–4980, Dec. 2013.
- [67] S. Y. Doerflinger, A. L. Throop, and M. M. Herbst-Kralovetz, “Bacteria in the Vaginal Microbiome Alter the Innate Immune Response and Barrier Properties of the Human Vaginal Epithelia in a Species-Specific Manner.,” *J INFECT DIS*, Feb. 2014.
- [68] E. V. Valore, D. J. Wiley, and T. Ganz, “Reversible deficiency of antimicrobial polypeptides in bacterial vaginosis.,” *Infection and Immunity*, vol. 74, no. 10, pp. 5693–5702, Oct. 2006.
- [69] L. Furci, F. Sironi, M. Tolazzi, L. Vassena, and P. Lusso, “Alpha-defensins block the early steps of HIV-1 infection: interference with the binding of gp120 to CD4.,” *Blood*, vol. 109, no. 7, pp. 2928–2935, Apr. 2007.
- [70] W. Wang, S. M. Owen, D. L. Rudolph, A. M. Cole, T. Hong, A. J. Waring, R. B. Lal, and R. I. Lehrer, “Activity of alpha- and theta-defensins against primary isolates of HIV-1.,” *J. Immunol.*, vol. 173, no. 1, pp. 515–520, Jul. 2004.
- [71] Z. Wu, F. Cocchi, D. Gentles, B. Ericksen, J. Lubkowski, A. Devico, R. I. Lehrer, and W. Lu, “Human neutrophil alpha-defensin 4 inhibits HIV-1 infection in vitro.,” *FEBS Lett.*, vol. 579, no. 1, pp. 162–166, Jan. 2005.
- [72] W. Wang, A. M. Cole, T. Hong, and A. J. Waring, “Retrocyclin, an antiretroviral θ -defensin, is a lectin,” *The Journal of Immunology*, 2003.
- [73] C. Münk, G. Wei, O. O. Yang, A. J. Waring, W. Wang, T. Hong, R. I. Lehrer, N. R. Landau, and A. M. Cole, “The theta-defensin, retrocyclin, inhibits HIV-1 entry.,” *AIDS Res. Hum. Retroviruses*, vol. 19, no. 10, pp. 875–881, Oct. 2003.
- [74] P. J. Maddon, A. G. Dalglish, J. S. McDougal, P. R. Clapham, R. A. Weiss, and R. Axel, “The T4 gene encodes the AIDS virus receptor and is

- expressed in the immune system and the brain,” *Cell*, vol. 47, no. 3, pp. 333–348, Jul. 1986.
- [75] D. Nasioudis, J. Beghini, A. M. Bongiovanni, P. C. Giraldo, I. M. Linhares, and S. S. Witkin, “ α -Amylase in Vaginal Fluid: Association With Conditions Favorable to Dominance of Lactobacillus.,” *Reprod Sci*, Apr. 2015.
- [76] T. Orfanelli, A. Jayaram, G. Doulaveris, L. J. Forney, W. J. Ledger, and S. S. Witkin, “Human Epididymis Protein 4 and Secretory Leukocyte Protease Inhibitor in Vaginal Fluid: Relation to Vaginal Components and Bacterial Composition,” *Reproductive Sciences*, vol. 21, no. 4, pp. 538–542, Mar. 2014.
- [77] B. J. Moncla, T. A. Mietzner, and S. L. Hillier, “In vitro activity of cationic peptides against Neisseria gonorrhoeae and vaginal Lactobacillus species: The effect of divalent cations,” *Advances in Bioscience and Biotechnology*, vol. 2012, no. 3, pp. 249–255, Jun. 2012.
- [78] E. V. Valore, C. H. Park, S. L. Igrati, and T. Ganz, “Antimicrobial components of vaginal fluid,” *American Journal of Obstetrics and Gynecology*, vol. 187, no. 3, pp. 561–568, Sep. 2002.
- [79] H. Pandit, S. Gopal, A. Sonawani, A. K. Yadav, A. S. Qaseem, H. Warke, A. Patil, R. Gajbhiye, V. Kulkarni, M. A. Al-Mozaini, S. Idicula-Thomas, U. Kishore, and T. Madan, “Surfactant Protein D Inhibits HIV-1 Infection of Target Cells via Interference with gp120-CD4 Interaction and Modulates Pro-Inflammatory Cytokine Production,” *PLoS ONE*, vol. 9, no. 7, pp. e102395–11, Jul. 2014.
- [80] G. D. Gaiha, T. Dong, N. Palaniyar, D. A. Mitchell, K. B. M. Reid, and H. W. Clark, “Surfactant protein A binds to HIV and inhibits direct infection of CD4+ cells, but enhances dendritic cell-mediated viral transfer.,” *The Journal of Immunology*, vol. 181, no. 1, pp. 601–609, Jul. 2008.
- [81] G. G. G. Donders, A. Vereecken, E. Bosmans, A. Dekeersmaecker, G. Salembier, and B. Spitz, “Definition of a type of abnormal vaginal flora that is distinct from bacterial vaginosis: aerobic vaginitis.,” *BJOG*, vol. 109, no. 1, pp. 34–43, Jan. 2002.
- [82] G. G. G. Donders, “Definition and classification of abnormal vaginal flora.,” *Best Pract Res Clin Obstet Gynaecol*, vol. 21, no. 3, pp. 355–373, Jun. 2007.
- [83] C. Han, W. Wu, A. Fan, Y. Wang, H. Zhang, Z. Chu, C. Wang, and F. Xue, “Diagnostic and therapeutic advancements for aerobic vaginitis.,” *Arch Gynecol Obstet*, Nov. 2014.
- [84] S. Cauci, “Vaginal Immunity in Bacterial Vaginosis.,” *Curr Infect Dis Rep*, vol. 6, no. 6, pp. 450–456, Nov. 2004.
- [85] S. S. Witkin, I. M. Linhares, P. Giraldo, and W. J. Ledger, “An altered immunity hypothesis for the development of symptomatic bacterial vaginosis.,” *Clin. Infect. Dis.*, vol. 44, no. 4, pp. 554–557, Feb. 2007.
- [86] S. Cauci, S. Guaschino, D. de Aloysio, S. Driussi, D. De Santo, P. Penacchioni, and F. Quadrifoglio, “Interrelationships of interleukin-8 with interleukin-1beta and neutrophils in vaginal fluid of healthy and bacterial

- vaginosis positive women.," *Mol. Hum. Reprod.*, vol. 9, no. 1, pp. 53–58, Jan. 2003.
- [87] E. K. Libby, K. E. Pascal, E. Mordechai, M. E. Adelson, and J. P. Trama, "Atopobium vaginae triggers an innate immune response in an in vitro model of bacterial vaginosis.," *Microbes Infect.*, vol. 10, no. 4, pp. 439–446, Apr. 2008.
- [88] A. N. Chaudry, P. J. Travers, J. Yuenger, L. Colletta, P. Evans, J. M. Zenilman, and A. Tummon, "Analysis of vaginal acetic acid in patients undergoing treatment for bacterial vaginosis.," *J. Clin. Microbiol.*, vol. 42, no. 11, pp. 5170–5175, Nov. 2004.
- [89] D. E. O'Hanlon, T. R. Moench, and R. A. Cone, "Vaginal pH and microbicidal lactic acid when lactobacilli dominate the microbiota.," *PLoS ONE*, vol. 8, no. 11, p. e80074, 2013.
- [90] P. Mirmonsef, M. R. Zariffard, D. Gilbert, H. Makinde, A. L. Landay, and G. T. Spear, "Short-Chain Fatty Acids Induce Pro-Inflammatory Cytokine Production Alone and in Combination with Toll-Like Receptor Ligands," *American Journal of Reproductive Immunology*, vol. 67, no. 5, pp. 391–400, Nov. 2011.
- [91] S. Al-Mushrif, A. Eley, and B. M. Jones, "Inhibition of chemotaxis by organic acids from anaerobes may prevent a purulent response in bacterial vaginosis.," *Journal of Medical Microbiology*, vol. 49, no. 11, pp. 1023–1030, Nov. 2000.
- [92] J. R. Schwebke, "New concepts in the etiology of bacterial vaginosis.," *Curr Infect Dis Rep*, vol. 11, no. 2, pp. 143–147, Mar. 2009.
- [93] E. R. Boskey, R. A. Cone, K. J. Whaley, and T. R. Moench, "Origins of vaginal acidity: high D/L lactate ratio is consistent with bacteria being the primary source.," *Hum Reprod*, vol. 16, no. 9, pp. 1809–1813, Sep. 2001.
- [94] Z. Gong, Y. Luna, P. Yu, and H. Fan, "Lactobacilli Inactivate Chlamydia trachomatis through Lactic Acid but Not H₂O₂.," vol. 9, no. 9, p. e107758, 2014.
- [95] C. Conti, C. Malacrino, and P. Mastromarino, "Inhibition of herpes simplex virus type 2 by vaginal lactobacilli.," *J. Physiol. Pharmacol.*, vol. 60, pp. 19–26, Dec. 2009.
- [96] C. E. Isaacs and W. Xu, "Theaflavin-3,3'-digallate and lactic acid combinations reduce herpes simplex virus infectivity.," *Antimicrob. Agents Chemother.*, vol. 57, no. 8, pp. 3806–3814, Aug. 2013.
- [97] M. Aldunate, D. Tyssen, A. Johnson, T. Zakir, S. Sonza, T. Moench, R. Cone, and G. Tachedjian, "Vaginal concentrations of lactic acid potentially inactivate HIV.," *J. Antimicrob. Chemother.*, vol. 68, no. 9, pp. 2015–2025, Sep. 2013.
- [98] M. Aldunate, D. Tyssen, C. Latham, P. Ramsland, P. Perlmutter, T. Moench, R. Cone, and G. Tachedjian, "Vaginal Concentrations of Lactic Acid Potentially Inactivate HIV-1 Compared to Short Chain Fatty Acids Present During Bacterial Vaginosis.," *AIDS Res. Hum. Retroviruses*, vol. 30, p. A228, Oct. 2014.
- [99] D. E. O'Hanlon, T. R. Moench, and R. A. Cone, "In vaginal fluid, bacteria

- associated with bacterial vaginosis can be suppressed with lactic acid but not hydrogen peroxide.," *BMC Infect. Dis.*, vol. 11, no. 1, p. 200, 2011.
- [100] A. Hearps, R. Gugasyan, D. Srbinovski, D. Tyssen, M. Aldunate, D. J. Anderson, R. Cone, and G. Tachedjian, "Lactic Acid, a Vaginal Microbiota Metabolite, Elicits an Anti-inflammatory Response from Vaginal and Cervical Epithelial Cells.," *AIDS Res. Hum. Retroviruses*, vol. 30, pp. A238–9, Oct. 2014.
- [101] S. S. Witkin, S. Alvi, A. M. Bongiovanni, I. M. Linhares, and W. J. Ledger, "Lactic acid stimulates interleukin-23 production by peripheral blood mononuclear cells exposed to bacterial lipopolysaccharide.," *FEMS Immunol. Med. Microbiol.*, vol. 61, no. 2, pp. 153–158, Mar. 2011.
- [102] H. Mossop, I. M. Linhares, A. M. Bongiovanni, W. J. Ledger, and S. S. Witkin, "Influence of Lactic Acid on Endogenous and Viral RNA-Induced Immune Mediator Production by Vaginal Epithelial Cells," *Obstet Gynecol*, vol. 118, no. 4, pp. 840–846, Oct. 2011.
- [103] T. Matsuki, T. Pédrón, B. Regnault, C. Mulet, T. Hara, and P. J. Sansonetti, "Epithelial Cell Proliferation Arrest Induced by Lactate and Acetate from *Lactobacillus casei* and *Bifidobacterium breve*," vol. 8, no. 4, p. e63053, Apr. 2013.
- [104] J. Beghini, I. Linhares, P. Giraldo, W. Ledger, and S. Witkin, "Differential expression of lactic acid isomers, extracellular matrix metalloproteinase inducer, and matrix metalloproteinase-8 in vaginal fluid from women with vaginal disorders.," *BJOG*, Sep. 2014.
- [105] S. S. Witkin, H. Mendes-Soares, I. M. Linhares, A. Jayaram, W. J. Ledger, and L. J. Forney, "Influence of Vaginal Bacteria and d- and l-Lactic Acid Isomers on Vaginal Extracellular Matrix Metalloproteinase Inducer: Implications for Protection against Upper Genital Tract Infections," *mBio*, vol. 4, no. 4, pp. e00460–13, 2013.
- [106] K. L. Nunn, Y.-Y. Wang, D. Harit, M. S. Humphrys, B. Ma, R. Cone, J. Ravel, and S. K. Lai, "Enhanced Trapping of HIV-1 by Human Cervicovaginal Mucus Is Associated with *Lactobacillus crispatus*-Dominant Microbiota.," *mBio*, vol. 6, no. 5, pp. e01084–15, 2015.
- [107] K. Nunn, Y.-Y. Wang, D. Harit, R. Cone, and S. Lai, "Influence of vaginal microbiota on the diffusional barrier properties of cervicovaginal mucus.," *AIDS Res. Hum. Retroviruses*, vol. 30, p. A234, Oct. 2014.
- [108] K. Arnold, K. Birse, L. Mckinnon, J. Lingappa, R. Novak, G. Westmacott, T. B. Ball, D. Lauffenburger, and A. Burgener, "Mucosal Integrity Factors Are Perturbed during Bacterial Vaginosis: A Proteomic Analysis.," *AIDS Res. Hum. Retroviruses*, vol. 30, p. A30, Oct. 2014.
- [109] W. G. Lewis, L. S. Robinson, J. Perry, J. L. Bick, J. F. Peipert, J. E. Allsworth, and A. L. Lewis, "Hydrolysis of secreted sialoglycoprotein immunoglobulin A (IgA) in ex vivo and biochemical models of bacterial vaginosis.," *J. Biol. Chem.*, vol. 287, no. 3, pp. 2079–2089, Jan. 2012.
- [110] W. G. Lewis, L. S. Robinson, N. M. Gilbert, J. C. Perry, and A. L. Lewis, "Degradation, foraging, and depletion of mucus sialoglycans by the vagina-adapted Actinobacterium *Gardnerella vaginalis*.," *J. Biol. Chem.*, vol. 288,

- no. 17, pp. 12067–12079, Apr. 2013.
- [111] M. Zilnyte, Č. Venclovas, A. Zvirbliene, and M. Pleckaityte, “The Cytolytic Activity of Vaginolysin Strictly Depends on Cholesterol and Is Potentiated by Human CD59,” *Toxins*, vol. 7, no. 1, pp. 110–128, Jan. 2015.
- [112] S. E. Gelber, J. L. Aguilar, K. L. T. Lewis, and A. J. Ratner, “Functional and phylogenetic characterization of Vaginolysin, the human-specific cytolysin from *Gardnerella vaginalis*,” *J. Bacteriol.*, vol. 190, no. 11, pp. 3896–3903, Jun. 2008.
- [113] B. J. Moncla, C. Chappell, B. M. Debo, I. S. Macio, K. E. Bunge, and S. L. Hillier, “The Effects of Hormones and Vaginal Microflora on the Content of MUC1, MUC4, MUC5AC and MUC7 in the Cervicovaginal Fluid (CVF),” *AIDS Res. Hum. Retroviruses*, vol. 30, p. A29, Oct. 2014.
- [114] N. van Oostrum, P. De Sutter, J. Meys, and H. Verstraelen, “Risks associated with bacterial vaginosis in infertility patients: a systematic review and meta-analysis,” *Hum Reprod*, vol. 28, no. 7, pp. 1809–1815, Jun. 2013.
- [115] G. G. Donders, K. Van Calsteren, G. Bellen, R. Reybrouck, T. Van den Bosch, I. Riphagen, and S. Van Lierde, “Predictive value for preterm birth of abnormal vaginal flora, bacterial vaginosis and aerobic vaginitis during the first trimester of pregnancy,” *BJOG*, vol. 116, no. 10, pp. 1315–1324, Jun. 2009.
- [116] D. B. Nelson, A. Hanlon, I. Nachamkin, C. Haggerty, D. S. Mastrogiannis, C. Liu, and D. N. Fredricks, “Early Pregnancy Changes in Bacterial Vaginosis-Associated Bacteria and Preterm Delivery,” *Paediatr Perinat Epidemiol*, vol. 28, pp. 88–96, Jan. 2014.
- [117] D. B. DiGiulio, B. J. Callahan, P. J. McMurdie, E. K. Costello, D. J. Lyell, A. Robaczewska, C. L. Sun, D. S. A. Goltsman, R. J. Wong, G. Shaw, D. K. Stevenson, S. P. Holmes, and D. A. Relman, “Temporal and spatial variation of the human microbiota during pregnancy,” *Proc. Natl. Acad. Sci. U.S.A.*, pp. 201502875–6, Aug. 2015.
- [118] A. Jakovljević, M. Bogovac, A. Nikolić, M. M. Tošić, Z. Novaković, and Z. Stajić, “The influence of bacterial vaginosis on gestational week of the completion of delivery and biochemical markers of inflammation in the serum,” *Vojnosanit Pregl*, vol. 71, no. 10, pp. 931–935, Oct. 2014.
- [119] R. Romero, S. S. Hassan, P. Gajer, A. L. Tarca, D. W. Fadrosch, J. Bieda, P. Chaemsathong, J. Miranda, T. Chaiworapongsa, and J. Ravel, “The vaginal microbiota of pregnant women who subsequently have spontaneous preterm labor and delivery and those with a normal delivery at term,” *Microbiome*, vol. 2, no. 1, p. 18, 2014.
- [120] L. M. Gómez, M. D. Sammel, D. H. Appleby, M. A. Elovitz, D. A. Baldwin, M. K. Jeffcoat, G. A. Macones, and S. Parry, “Evidence of a gene-environment interaction that predisposes to spontaneous preterm birth: a role for asymptomatic bacterial vaginosis and DNA variants in genes that control the inflammatory response,” *American Journal of Obstetrics and Gynecology*, vol. 202, no. 4, pp. 386.e1–6, Apr. 2010.

- [121] N. M. Jones, C. Holzman, K. H. Friderici, K. Jernigan, H. Chung, J. Wirth, and R. Fisher, "Interplay of cytokine polymorphisms and bacterial vaginosis in the etiology of preterm delivery.," *Journal of Reproductive Immunology*, vol. 87, no. 1, pp. 82–89, Dec. 2010.
- [122] R. L. Goldenberg, J. C. Hauth, and W. W. Andrews, "Intrauterine infection and preterm delivery.," *N. Engl. J. Med.*, vol. 342, no. 20, pp. 1500–1507, May 2000.
- [123] S. S. Witkin, "The vaginal microbiome, vaginal anti-microbial defence mechanisms and the clinical challenge of reducing infection-related preterm birth," *BJOG*, vol. 122, pp. 213–218, Oct. 2014.
- [124] M. G. Gravett, S. S. Witkin, G. J. Haluska, J. L. Edwards, M. J. Cook, and M. J. Novy, "An experimental model for intraamniotic infection and preterm labor in rhesus monkeys.," *YMOB*, vol. 171, no. 6, pp. 1660–1667, Dec. 1994.
- [125] A. E. King, N. Wheelhouse, S. Cameron, S. E. McDonald, K.-F. Lee, G. Entrican, H. O. D. Critchley, and A. W. Horne, "Expression of secretory leukocyte protease inhibitor and elafin in human fallopian tube and in an in-vitro model of Chlamydia trachomatis infection.," *Hum Reprod*, vol. 24, no. 3, pp. 679–686, Mar. 2009.
- [126] C. R. Eade, C. Diaz, M. P. Wood, K. Anastos, B. K. Patterson, P. Gupta, A. L. Cole, and A. M. Cole, "Identification and characterization of bacterial vaginosis-associated pathogens using a comprehensive cervical-vaginal epithelial coculture assay.," *PLoS ONE*, vol. 7, no. 11, p. e50106, 2012.
- [127] M. D. Cooper, M. H. Roberts, O. L. Barauskas, and G. A. Jarvis, "Secretory leukocyte protease inhibitor binds to Neisseria gonorrhoeae outer membrane opacity protein and is bactericidal.," *Am. J. Reprod. Immunol.*, vol. 68, no. 2, pp. 116–127, Aug. 2012.
- [128] L. M. Hafner, "Pathogenesis of fallopian tube damage caused by Chlamydia trachomatis infections," *Contraception*, vol. 92, no. 2, pp. 108–115, Aug. 2015.
- [129] R. E. Barlow, I. D. Cooke, O. Odukoya, M. K. Heatley, J. Jenkins, G. Narayansingh, S. S. Ramsewak, and A. Eley, "The prevalence of Chlamydia trachomatis in fresh tissue specimens from patients with ectopic pregnancy or tubal factor infertility as determined by PCR and in-situ hybridisation.," *Journal of Medical Microbiology*, vol. 50, no. 10, pp. 902–908, Oct. 2001.
- [130] V. L. Tyutyunnik, N. E. Kan, N. A. Lomova, T. E. Karapetyan, E. A. Kogan, and A. I. Shchyogolev, "Role of Innate Immunity in Pregnant Patients with Vulvovaginal Infections in the Development of Intrauterine Infection in the Newborn," *Bull. Exp. Biol. Med.*, vol. 158, no. 1, pp. 74–76, Nov. 2014.
- [131] S. Cauci and J. F. Culhane, "Modulation of vaginal immune response among pregnant women with bacterial vaginosis by Trichomonas vaginalis, Chlamydia trachomatis, Neisseria gonorrhoeae, and yeast," *American Journal of Obstetrics and Gynecology*, vol. 196, no. 2, pp. 133.e1–133.e7, Feb. 2007.

- [132] R. Cruicksank and A. Sharman, *The biology of the vagina in the human subject II. The bacterial flora and secretion of the vagina at various age periods and their relation to glycogen in the ...* Journal of Obstetrics and Gynaecology British Empire, 1934.
- [133] R. M. Brotman, M. D. Shardell, P. Gajer, D. Fadrosch, K. Chang, M. I. Silver, R. P. Viscidi, A. E. Burke, J. Ravel, and P. E. Gravitt, "Association between the vaginal microbiota, menopause status, and signs of vulvovaginal atrophy.," *Menopause*, Sep. 2013.
- [134] M. R. Hammerschlag, S. Alpert, I. Rosner, P. Thurston, D. Semine, D. McComb, and W. M. McCormack, "Microbiology of the vagina in children: normal and potentially pathogenic organisms.," *Pediatrics*, vol. 62, no. 1, pp. 57–62, Jul. 1978.
- [135] C. L. Galhardo, J. M. Soares, R. S. Simões, M. A. Haidar, G. Rodrigues de Lima, and E. C. Baracat, "Estrogen effects on the vaginal pH, flora and cytology in late postmenopause after a long period without hormone therapy.," *Clin Exp Obstet Gynecol*, vol. 33, no. 2, pp. 85–89, 2006.
- [136] S. L. Hillier and R. J. Lau, "Vaginal microflora in postmenopausal women who have not received estrogen replacement therapy," *Clinical infectious diseases*, vol. 25, no. 2, pp. S123–S126, 1997.
- [137] M. R. Hammerschlag, S. Alpert, A. B. Onderdonk, P. Thurston, E. Drude, W. M. McCormack, and J. G. Bartlett, "Anaerobic microflora of the vagina in children.," *YMOB*, vol. 131, no. 8, pp. 853–856, Aug. 1978.
- [138] K. Nilsson, B. Risberg, and G. Heimer, "The vaginal epithelium in the postmenopause--cytology, histology and pH as methods of assessment.," *Maturitas*, vol. 21, no. 1, pp. 51–56, Jan. 1995.
- [139] R. Cruickshank and A. Sharman, "The biology of the vagina in the human subject," *BJOG: An International Journal of Obstetrics & Gynaecology*, 1934.
- [140] R. Cruickshank, "The conversion of the glycogen of the vagina into lactic acid," *The Journal of Pathology and Bacteriology*, vol. 39, no. 1, pp. 213–219, Jul. 1934.
- [141] D. E. Stewart-Tull, "Evidence that Vaginal Lactobacilli do not Ferment Glycogen," *YMOB*, vol. 88, pp. 676–679, Mar. 1964.
- [142] G. T. Spear, A. L. French, D. Gilbert, M. R. Zariffard, P. Mirmonsef, T. H. Sullivan, W. W. Spear, A. Landay, S. Micci, B.-H. Lee, and B. R. Hamaker, "Human α -amylase present in lower-genital-tract mucosal fluid processes glycogen to support vaginal colonization by Lactobacillus.," *J INFECT DIS*, vol. 210, no. 7, pp. 1019–1028, Oct. 2014.
- [143] J. C. Bernbaum, D. M. Umbach, N. B. Ragan, J. L. Ballard, J. I. Archer, H. Schmidt-Davis, and W. J. Rogan, "Pilot Studies of Estrogen-Related Physical Findings in Infants," *Environ. Health Perspect.*, vol. 116, no. 3, pp. 416–420, Dec. 2007.
- [144] W. Filipowicz, S. N. Bhattacharyya, and N. Sonenberg, "Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?," *Nature Reviews Genetics*, vol. 9, no. 2, pp. 102–114, Feb. 2008.
- [145] L. A. O'Neill, F. J. Sheedy, and C. E. McCoy, "MicroRNAs: the fine-tuners

- of Toll-like receptor signalling.," *Nat Rev Immunol*, vol. 11, no. 3, pp. 163–175, Mar. 2011.
- [146] M. Ha and V. N. Kim, "Regulation of microRNA biogenesis," *Nat. Rev. Mol. Cell Biol.*, vol. 15, no. 8, pp. 509–524, Aug. 2014.
- [147] V. N. Kim, J. Han, and M. C. Siomi, "Biogenesis of small RNAs in animals.," *Nat. Rev. Mol. Cell Biol.*, vol. 10, no. 2, pp. 126–139, Feb. 2009.
- [148] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic Acids Research*, vol. 42, no. 1, pp. D68–D73, Dec. 2013.
- [149] Y. Lee, M. Kim, J. Han, K.-H. Yeom, S. Lee, S. H. Baek, and V. N. Kim, "MicroRNA genes are transcribed by RNA polymerase II.," *EMBO J.*, vol. 23, no. 20, pp. 4051–4060, Oct. 2004.
- [150] X. Cai, C. H. Hagedorn, and B. R. Cullen, "Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs.," *RNA*, vol. 10, no. 12, pp. 1957–1966, Dec. 2004.
- [151] A. Rodriguez, S. GRIFFITHS-JONES, J. L. Ashurst, and A. Bradley, "Identification of mammalian microRNA host genes and transcription units.," *Genome Research*, vol. 14, no. 10, pp. 1902–1910, Oct. 2004.
- [152] Y.-S. Long, G.-F. Deng, X.-S. Sun, Y.-H. Yi, T. Su, Q.-H. Zhao, and W.-P. Liao, "Identification of the transcriptional promoters in the proximal regions of human microRNA genes.," *Mol. Biol. Rep.*, vol. 38, no. 6, pp. 4153–4157, Aug. 2011.
- [153] Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Rådmark, S. Kim, and V. N. Kim, "The nuclear RNase III Drosha initiates microRNA processing.," *Nature*, vol. 425, no. 6956, pp. 415–419, Sep. 2003.
- [154] E. Lund, S. Güttinger, A. Calado, J. E. Dahlberg, and U. Kutay, "Nuclear export of microRNA precursors.," *Science*, vol. 303, no. 5654, pp. 95–98, Jan. 2004.
- [155] R. Yi, Y. Qin, I. G. Macara, and B. R. Cullen, "Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs.," *Genes & Development*, vol. 17, no. 24, pp. 3011–3016, Dec. 2003.
- [156] M. T. Bohnsack, K. Czaplinski, and D. Gorlich, "Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs.," *RNA*, vol. 10, no. 2, pp. 185–191, Feb. 2004.
- [157] I. J. Macrae, K. Zhou, F. Li, A. Repic, A. N. Brooks, W. Z. Cande, P. D. Adams, and J. A. Doudna, "Structural basis for double-stranded RNA processing by Dicer.," *Science*, vol. 311, no. 5758, pp. 195–198, Jan. 2006.
- [158] J.-E. Park, I. Heo, Y. Tian, D. K. Simanshu, H. Chang, D. Jee, D. J. Patel, and V. N. Kim, "Dicer recognizes the 5' end of RNA for efficient and accurate processing.," *Nature*, vol. 475, no. 7355, pp. 201–205, Jul. 2011.
- [159] A. Eulalio, F. Tritschler, and E. Izaurralde, "The GW182 protein family in animal cells: new insights into domains required for miRNA-mediated gene silencing.," *RNA*, vol. 15, no. 8, pp. 1433–1442, Aug. 2009.
- [160] G. Meister, M. Landthaler, A. Patkaniowska, Y. Dorsett, G. Teng, and T. TUSCHL, "Human Argonaute2 Mediates RNA Cleavage Targeted by

- miRNAs and siRNAs,” *Mol. Cell*, vol. 15, no. 2, pp. 185–197, Jul. 2004.
- [161] D. Baillat and R. Shiekhattar, “Functional dissection of the human TNRC6 (GW182-related) family of proteins,” *Molecular and Cellular Biology*, vol. 29, no. 15, pp. 4144–4155, Aug. 2009.
- [162] C. L. Noland and J. A. Doudna, “Multiple sensors ensure guide strand selection in human RNAi pathways,” *RNA*, vol. 19, no. 5, pp. 639–648, May 2013.
- [163] H. Su, M. I. Trombly, J. Chen, and X. Wang, “Essential and overlapping functions for mammalian Argonautes in microRNA silencing,” *Genes & Development*, vol. 23, no. 3, pp. 304–317, Feb. 2009.
- [164] M. Livingstone, E. Atas, A. Meller, and N. Sonenberg, “Mechanisms governing the control of mRNA translation,” *Phys Biol*, vol. 7, no. 2, p. 021001, May 2010.
- [165] S. E. Wells, P. E. Hillner, R. D. Vale, and A. B. Sachs, “Circularization of mRNA by eukaryotic translation initiation factors,” *Mol. Cell*, vol. 2, no. 1, pp. 135–140, Jul. 1998.
- [166] R. J. Jackson, C. U. T. Hellen, and T. V. Pestova, “The mechanism of eukaryotic translation initiation and principles of its regulation,” *Nat. Rev. Mol. Cell Biol.*, vol. 11, no. 2, pp. 113–127, Feb. 2010.
- [167] J. Béthune, C. G. Artus-Revel, and W. Filipowicz, “Kinetic analysis reveals successive steps leading to miRNA-mediated silencing in mammalian cells,” *Nature Publishing Group*, vol. 13, no. 8, pp. 716–723, Jun. 2012.
- [168] H. A. Meijer, Y. W. Kong, W. T. Lu, A. Wilczynska, R. V. Spriggs, S. W. Robinson, J. D. Godfrey, A. E. Willis, and M. Bushell, “Translational repression and eIF4A2 activity are critical for microRNA-mediated gene regulation,” *Science*, vol. 340, no. 6128, pp. 82–85, Apr. 2013.
- [169] G. Mathonnet, M. R. Fabian, Y. V. Svitkin, A. Parsyan, L. Huck, T. Murata, S. Biffo, W. C. Merrick, E. Darzynkiewicz, R. S. Pillai, W. Filipowicz, T. F. Duchaine, and N. Sonenberg, “MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F,” *Science*, vol. 317, no. 5845, pp. 1764–1767, Sep. 2007.
- [170] C. P. Petersen, M.-E. Bordeleau, J. Pelletier, and P. A. Sharp, “Short RNAs Repress Translation after Initiation in Mammalian Cells,” *Mol. Cell*, vol. 21, no. 4, pp. 533–542, Feb. 2006.
- [171] T. T. Tat, P. A. Maroney, S. Chamnongpol, and J. Collier, “Cotranslational microRNA mediated messenger RNA destabilization,” *eLife*, 2016.
- [172] S. W. Eichhorn, H. Guo, S. E. McGeary, R. A. Rodriguez-Mias, C. Shin, D. Baek, S.-H. Hsu, K. Ghoshal, J. Villén, and D. P. Bartel, “mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues,” *Mol. Cell*, vol. 56, no. 1, pp. 104–115, Oct. 2014.
- [173] L. Wu, J. Fan, and J. G. Belasco, “MicroRNAs direct rapid deadenylation of mRNA,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, no. 11, pp. 4034–4039, Mar. 2006.
- [174] A. Fukao, Y. Mishima, N. Takizawa, S. Oka, H. Imataka, J. Pelletier, N. Sonenberg, C. Thoma, and T. Fujiwara, “MicroRNAs Trigger Dissociation

- of eIF4AI and eIF4AII from Target mRNAs in Humans,” *Mol. Cell*, vol. 56, no. 1, pp. 79–89, Oct. 2014.
- [175] E. P. Ricci, T. Limousin, R. Soto-Rifo, P. S. Rubilar, D. Decimo, and T. Ohlmann, “miRNA repression of translation in vitro takes place during 43S ribosomal scanning,” *Nucleic Acids Research*, vol. 41, no. 1, pp. 586–598, Dec. 2012.
- [176] R. S. Pillai, S. N. Bhattacharyya, C. G. Artus, T. Zoller, N. Cougot, E. Basyuk, E. Bertrand, and W. Filipowicz, “Inhibition of translational initiation by Let-7 MicroRNA in human cells,” *Science*, vol. 309, no. 5740, pp. 1573–1576, Sep. 2005.
- [177] M. Kiriakidou, G. S. Tan, S. Lamprinaki, M. De Planell-Saguer, P. T. Nelson, and Z. Mourelatos, “An mRNA m7G cap binding-like motif within human Ago2 represses translation,” *Cell*, vol. 129, no. 6, pp. 1141–1151, Jun. 2007.
- [178] T. P. Chendrimada, K. J. Finn, X. Ji, D. Baillat, R. I. Gregory, S. A. Liebhaber, A. E. Pasquinelli, and R. Shiekhattar, “MicroRNA silencing through RISC recruitment of eIF6,” *Nature*, vol. 447, no. 7146, pp. 823–828, May 2007.
- [179] H. Mathys, J. Basquin, S. Ozgur, M. Czarnocki-Cieciura, F. Bonneau, A. Aartse, A. Dziembowski, M. Nowotny, E. Conti, and W. Filipowicz, “Structural and Biochemical Insights to the Role of the CCR4-NOT Complex and DDX6 ATPase in MicroRNA Repression,” *Mol. Cell*, vol. 54, no. 5, pp. 751–765, Jun. 2014.
- [180] Y. Chen, A. Boland, D. Kuzuoğlu-Öztürk, P. Bawankar, B. Loh, C.-T. Chang, O. Weichenrieder, and E. Izaurralde, “A DDX6-CNOT1 Complex and W-Binding Pockets in CNOT9 Reveal Direct Links between miRNA Target Recognition and Silencing,” *Mol. Cell*, vol. 54, no. 5, pp. 737–750, Jun. 2014.
- [181] J. E. Braun, E. Huntzinger, M. Fauser, and E. Izaurralde, “GW182 Proteins Directly Recruit Cytoplasmic Deadenylation Complexes to miRNA Targets,” *Mol. Cell*, vol. 44, no. 1, pp. 120–133, Oct. 2011.
- [182] M. R. Fabian, M. K. Cieplak, F. Frank, M. Morita, J. Green, T. Srikumar, B. Nagar, T. Yamamoto, B. Raught, T. F. Duchaine, and N. Sonenberg, “miRNA-mediated deadenylation is orchestrated by GW182 through two conserved motifs that interact with CCR4-NOT,” *Nat Struct Mol Biol*, vol. 18, no. 11, pp. 1211–1217, Oct. 2011.
- [183] L. Zekri, D. K. G. L.-& Z. U. rk, and E. Izaurralde, “GW182 proteins cause PABP dissociation from silenced miRNA targets in the absence of deadenylation,” *EMBO J.*, vol. 32, no. 7, pp. 1052–1065, Mar. 2013.
- [184] T. Nishihara, L. Zekri, J. E. Braun, and E. Izaurralde, “miRISC recruits decapping factors to miRNA targets to enhance their degradation,” *Nucleic Acids Research*, vol. 41, no. 18, pp. 8692–8705, Oct. 2013.
- [185] T. Nishimura and M. R. Fabian, “Scanning for a unified model for translational repression by microRNAs,” *EMBO J.*, vol. 35, no. 11, pp. 1158–1159, May 2016.
- [186] M. M. W. Chong, G. Zhang, S. Cheloufi, T. A. Neubert, G. J. Hannon, and

- D. R. Littman, “Canonical and alternate functions of the microRNA biogenesis machinery,” *Genes & Development*, vol. 24, no. 17, pp. 1951–1960, Sep. 2010.
- [187] J. E. Babiarz, J. G. Ruby, Y. Wang, D. P. Bartel, and R. Blelloch, “Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs.,” *Genes & Development*, vol. 22, no. 20, pp. 2773–2785, Oct. 2008.
- [188] I. Heo, M. Ha, J. Lim, M.-J. Yoon, J.-E. Park, S. C. Kwon, H. Chang, and V. N. Kim, “Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs.,” *Cell*, vol. 151, no. 3, pp. 521–532, Oct. 2012.
- [189] Y.-K. Kim, B. Kim, and V. N. Kim, “Re-evaluation of the roles of DROSHA, Exportin 5, and DICER1 in microRNA biogenesis,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 113, no. 13, pp. E1881–E1889, Mar. 2016.
- [190] J.-S. Yang, T. Maurin, N. Robine, K. D. Rasmussen, K. L. Jeffrey, R. Chandwani, E. P. Papapetrou, M. Sadelain, D. O’Carroll, and E. C. Lai, “Conserved vertebrate mir-451 provides a platform for Dicer-independent, Ago2-mediated microRNA biogenesis.,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 34, pp. 15163–15168, Aug. 2010.
- [191] G. Ramaswami, R. Zhang, R. Piskol, L. P. Keegan, P. Deng, M. A. O’Connell, and J. B. Li, “Identifying RNA editing sites using RNA sequencing data alone,” *Nat. Methods*, vol. 10, no. 2, pp. 128–132, Jan. 2013.
- [192] L. Guo and F. Chen, “A challenge for miRNA: multiple isomiRs in miRNAomics.,” *Gene*, Apr. 2014.
- [193] B. A. Duerkop, S. Vaishnava, and L. V. Hooper, “Immune responses to the microbiota at the intestinal mucosal surface.,” *Immunity*, vol. 31, no. 3, pp. 368–376, Sep. 2009.
- [194] P. Mirmonsef, D. Gilbert, M. R. Zariffard, B. R. Hamaker, A. Kaur, A. L. Landay, and G. T. Spear, “The Effects of Commensal Bacteria on Innate Immune Responses in the Female Genital Tract,” *American Journal of Reproductive Immunology*, vol. 65, no. 3, pp. 190–195, Dec. 2010.
- [195] S. Jia, H. Zhai, and M. Zhao, “MicroRNAs regulate immune system via multiple targets.,” *Discovery Medicine*, vol. 18, no. 100, pp. 237–247, Nov. 2014.
- [196] H.-M. Lee, D. T. Nguyen, and L.-F. Lu, “Progress and challenge of microRNA research in immunity.,” *Front Genet*, vol. 5, p. 178, 2014.
- [197] R. P. Singh, I. Massachi, S. Manickavel, S. Singh, N. P. Rao, S. Hasan, D. K. Mc Curdy, S. Sharma, D. Wong, B. H. Hahn, and H. Rehim, “The role of miRNA in inflammation and autoimmunity.,” *Autoimmun Rev*, Jul. 2013.
- [198] M. Duval, P. Cossart, and A. Lebreton, “Mammalian microRNAs and long noncoding RNAs in the host-bacterial pathogen crosstalk,” *Seminars in Cell & Developmental Biology*, pp. 1–34, Jul. 2016.
- [199] C. Staedel and F. Darfeuille, “MicroRNAs and bacterial infection,” vol. 15, no. 9, pp. 1496–1507, Jul. 2013.
- [200] C. Maudet, M. Mano, and A. Eulalio, “MicroRNAs in the interaction

- between host and bacterial pathogens,” *FEBS Lett.*, vol. 588, no. 22, pp. 4140–4147, Nov. 2014.
- [201] G. Dalmaso, H. T. T. Nguyen, Y. Yan, H. Laroui, M. A. Charania, S. Ayyadurai, S. V. Sitaraman, and D. Merlin, “Microbiota modulate host gene expression via microRNAs,” vol. 6, no. 4, p. e19293, 2011.
- [202] S. Liu, A. P. da Cunha, R. M. Rezende, R. Cialic, Z. Wei, L. Bry, L. E. Comstock, R. Gandhi, and H. L. Weiner, “The Host Shapes the Gut Microbiota via Fecal MicroRNA,” *Cell Host Microbe*, vol. 19, no. 1, pp. 32–43, Jan. 2016.
- [203] M. Biton, A. Levin, M. Slyper, I. Alkalay, E. Horwitz, H. Mor, S. Kredor-Russo, T. Avnit-Sagi, G. Cojocar, F. Zreik, Z. Bentwich, M. N. Poy, D. Artis, M. D. Walker, E. Hornstein, E. Pikarsky, and Y. Ben-Neriah, “Epithelial microRNAs regulate gut mucosal immunity via epithelium-T cell crosstalk,” *Nat Immunol*, vol. 12, no. 3, pp. 239–246, Mar. 2011.
- [204] N. H. Shahrin, S. Diakiw, L. A. Dent, A. L. Brown, and R. J. D’Andrea, “Conditional knockout mice demonstrate function of Klf5 as a myeloid transcription factor,” *Blood*, vol. 128, no. 1, pp. 55–59, Jul. 2016.
- [205] R. L. Baldwin, VI, S. Wu, W. Li, C. Li, B. J. Bequette, and R. W. Li, “Quantification of Transcriptome Responses of the Rumen Epithelium to Butyrate Infusion using RNA-seq Technology,” *Gene Regulation and Systems Biology*, vol. 6, pp. 67–80, 2012.
- [206] C.-J. Li, R. W. Li, R. L. Baldwin, VI, Le Ann Blomberg, S. Wu, and W. Li, “Transcriptomic Sequencing Reveals a Set of Unique Genes Activated by Butyrate-Induced Histone Modification,” *Gene Regulation and Systems Biology*, vol. 10, pp. 1–8, 2016.
- [207] S. Wu, R. W. Li, W. Li, and C.-J. Li, “Transcriptome Characterization by RNA-seq Unravels the Mechanisms of Butyrate-Induced Epigenomic Regulation in Bovine Cells,” *PLoS ONE*, vol. 7, no. 5, pp. e36940–15, May 2012.
- [208] “Multiple roles of class I HDACs in proliferation, differentiation, and development,” *Cellular and Molecular Life Sciences*, vol. 69, no. 13, pp. 2173–2187, Jul. 2012.
- [209] N. Turgeon, M. Blais, J.-M. Gagné, V. Tardif, F. Boudreau, N. Perreault, and C. Asselin, “HDAC1 and HDAC2 Restrain the Intestinal Inflammatory Response by Regulating Intestinal Epithelial Cell Differentiation,” *PLoS ONE*, vol. 8, no. 9, p. e73785, Sep. 2013.
- [210] S. Hu, T. S. Dong, S. R. Dalal, F. Wu, M. Bissonnette, J. H. Kwon, and E. B. Chang, “The Microbe-Derived Short Chain Fatty Acid Butyrate Targets miRNA-Dependent p21 Gene Expression in Human Colon Cancer,” *PLoS ONE*, vol. 6, no. 1, p. e16221, Jan. 2011.
- [211] T. Matsuki, T. Pédrón, B. Regnault, C. Mulet, T. Hara, and P. J. Sansonetti, “Epithelial cell proliferation arrest induced by lactate and acetate from *Lactobacillus casei* and *Bifidobacterium breve*,” *PLoS ONE*, vol. 8, no. 4, p. e63053, 2013.
- [212] Y. D. Bhutia, E. Babu, S. Ramachandran, S. Yang, M. Thangaraju, and V. Ganapathy, “SLC transporters as a novel class of tumour suppressors:

- identity, function and molecular mechanisms,” *Biochemical Journal*, vol. 473, no. 9, pp. 1113–1124, May 2016.
- [213] M. Thangaraju, G. Cresci, S. Itagaki, J. Mellinger, D. D. Browning, F. G. Berger, P. D. Prasad, and V. Ganapathy, “Sodium-Coupled Transport of the Short Chain Fatty Acid Butyrate by SLC5A8 and Its Relevance to Colon Cancer,” *J Gastrointest Surg*, vol. 12, no. 10, pp. 1773–1782, 2008.
- [214] M. A. Cox, J. Jackson, M. Stanton, A. Rojas-Triana, L. Bober, M. Laverty, X. Yang, F. Zhu, J. Liu, S. Wang, F. Monsma, G. Vassileva, M. Maguire, E. Gustafson, M. Bayne, C.-C. Chou, D. Lundell, and C.-H. Jenh, “Short-chain fatty acids act as antiinflammatory mediators by regulating prostaglandin E(2) and cytokines.,” *World J. Gastroenterol.*, vol. 15, no. 44, pp. 5549–5557, Nov. 2009.
- [215] J.-P. Segain, D. R. de la Blétière, A. Bourreille, V. Leray, N. Gervois, C. Rosales, L. Ferrier, C. Bonnet, H. M. Blottière, and J.-P. Galmiche, “Butyrate inhibits inflammatory responses through NFκB inhibition: implications for Crohn's disease,” *Gut*, vol. 47, no. 3, pp. 397–403, Sep. 2000.
- [216] H. Lührs, T. Gerke, J. G. Müller, R. Melcher, J. Schaubert, F. Boxberger, W. Scheppach, and T. Menzel, “Butyrate Inhibits NF-κB Activation in Lamina Propria Macrophages of Patients with Ulcerative Colitis,” *Scandinavian Journal of Gastroenterology*, vol. 37, no. 4, pp. 458–466, Jul. 2009.
- [217] D. C. Rio, M. Ares Jr, G. J. Hannon, and T. W. Nilsen, *RNA: a laboratory manual*. Cold Spring Harbor, 2011.
- [218] A. Schroeder, O. Mueller, S. Stocker, R. Salowsky, M. Leiber, M. Gassmann, S. Lightfoot, W. Menzel, M. Granzow, and T. Ragg, “The RIN: an RNA integrity number for assigning integrity values to RNA measurements,” *BMC Mol Biol*, vol. 7, no. 1, p. 3, 2006.
- [219] G. L. Mutter, D. Zahrieh, C. Liu, D. Neuberger, D. Finkelstein, H. E. Baker, and J. A. Warrington, “Comparison of frozen and RNALater solid tissue storage methods for use in RNA expression microarrays,” *BMC Genomics*, vol. 5, no. 1, p. 88, Nov. 2004.
- [220] S. Shore, J. M. Henderson, A. Lebedev, M. P. Salcedo, G. Zon, A. P. McCaffrey, N. Paul, and R. I. Hogrefe, “Small RNA Library Preparation Method for Next-Generation Sequencing Using Chemical Modifications to Prevent Adapter Dimer Formation,” *PLoS ONE*, vol. 11, no. 11, pp. e0167009–26, Nov. 2016.
- [221] S. Cirera and P. K. Busk, “Quantification of miRNAs by a simple and specific qPCR method.,” *Methods Mol. Biol.*, vol. 1182, no. 7, pp. 73–81, 2014.
- [222] F. Zeka, P. Mestdagh, and J. Vandesompele, “RT-qPCR-Based Quantification of Small Non-Coding RNAs.,” *Methods Mol. Biol.*, vol. 1296, no. 9, pp. 85–102, 2015.
- [223] C. Chen, D. A. Ridzon, A. J. Broomer, Z. Zhou, D. H. Lee, J. T. Nguyen, M. Barbisin, N. L. Xu, V. R. Mahuvakar, M. R. Andersen, K. Q. Lao, K. J. Livak, and K. J. Guegler, “Real-time quantification of microRNAs by stem-

- loop RT-PCR.,” *Nucleic Acids Research*, vol. 33, no. 20, pp. e179–e179, Nov. 2005.
- [224] V. Benes and M. Castoldi, “Expression profiling of microRNA using real-time quantitative PCR, how to use it and what is available.,” *Methods*, vol. 50, no. 4, pp. 244–249, Apr. 2010.
- [225] W. Zhao, X. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, and C. M. Perou, “Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling.,” *BMC Genomics*, vol. 15, no. 1, p. 419, Jun. 2014.
- [226] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics.,” *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, Jan. 2009.
- [227] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows–Wheeler transform,” *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009.
- [228] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.,” *Nat Protoc*, vol. 7, no. 3, pp. 562–578, Mar. 2012.
- [229] S. Tam, M.-S. Tsao, and J. D. McPherson, “Optimization of miRNA-seq data preprocessing.,” *Brief. Bioinformatics*, p. bbv019, Apr. 2015.
- [230] S. Anders, P. T. Pyl, and W. Huber, “HTSeq—a Python framework to work with high-throughput sequencing data.,” *Bioinformatics*, vol. 31, no. 2, pp. 166–169, Jan. 2015.
- [231] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, “RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.,” *Genome Research*, vol. 18, no. 9, pp. 1509–1517, Sep. 2008.
- [232] M. D. Robinson and G. K. Smyth, “Moderated statistical tests for assessing differences in tag abundance.,” *Bioinformatics*, vol. 23, no. 21, pp. 2881–2887, Nov. 2007.
- [233] D. J. McCarthy, Y. Chen, and G. K. Smyth, “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.,” *Nucleic Acids Research*, vol. 40, no. 10, pp. 4288–4297, May 2012.
- [234] A. L. Oberg, B. M. Bot, D. E. Grill, G. A. Poland, and T. M. Therneau, “Technical and biological variance structure in mRNA-Seq data: life in the real world.,” *BMC Genomics*, vol. 13, no. 1, p. 304, Jul. 2012.
- [235] G. Mi, Y. Di, and D. W. Schafer, “Goodness-of-fit tests and model diagnostics for negative binomial regression of RNA sequencing data.,” *PLoS ONE*, vol. 10, no. 3, p. e0119254, 2015.
- [236] J. Lu, J. K. Tomfohr, and T. B. Kepler, “Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach,” *BMC Bioinformatics*, vol. 6, no. 1, p. 165, Jun. 2005.
- [237] J. M. Hilbe, *Negative Binomial Regression*. Cambridge University Press, 2011.

- [238] S. Anders and W. Huber, “Differential expression analysis for sequence count data.,” *Genome biology*, vol. 11, no. 10, p. R106, 2010.
- [239] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2,” *bioRxiv*, 2014.
- [240] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.,” *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010.
- [241] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, “Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments,” *BMC Bioinformatics*, vol. 11, no. 1, p. 94, 2010.
- [242] M. D. Robinson and A. Oshlack, “A scaling normalization method for differential expression analysis of RNA-seq data.,” *Genome biology*, vol. 11, no. 3, p. R25, 2010.
- [243] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézic, French StatOmique Consortium, “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.,” *Brief. Bioinformatics*, vol. 14, no. 6, pp. 671–683, Nov. 2013.
- [244] M. W. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics,” *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, Jun. 2015.
- [245] G. K. Smyth, “Linear models and empirical bayes methods for assessing differential expression in microarray experiments.,” *Stat Appl Genet Mol Biol*, vol. 3, p. Article3, 2004.
- [246] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. Springer Science & Business Media, 2001.
- [247] M. D. Robinson and G. K. Smyth, “Small-sample estimation of negative binomial dispersion, with applications to SAGE data.,” *Biostatistics*, vol. 9, no. 2, pp. 321–332, Apr. 2008.
- [248] W. S. Noble, “How does multiple testing correction work?,” *Nat. Biotechnol.*, vol. 27, no. 12, pp. 1135–1137, Dec. 2009.
- [249] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel, “Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.,” *Genome biology*, vol. 14, no. 9, p. R95, Sep. 2013.
- [250] V. Jonsson, T. Österlund, O. Nerman, and E. Kristiansson, “Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics.,” *BMC Genomics*, vol. 17, no. 1, p. 78, Jan. 2016.
- [251] M. Stone, “Cross-validatory choice and assessment of statistical predictions,” *Journal of the royal statistical society Series B (...)*, 1974.
- [252] A. L. Boulesteix, S. Janitza, and J. Kruppa, “Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics,” *Data Mining and Knowledge Discovery*, 2012.
- [253] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification*

- and regression trees Belmont*. CA: Wadsworth International Group, 1984.
- [254] B. A. Goldstein, E. C. Polley, and F. B. S. Briggs, “Random forests for genetic association studies.,” *Stat Appl Genet Mol Biol*, vol. 10, no. 1, p. 32, 2011.
- [255] L. Breiman, “Bagging predictors,” *Mach Learn*, 1996.
- [256] W. G. Touw, J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and S. A. F. T. van Hijum, “Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?,” *Brief. Bioinformatics*, vol. 14, no. 3, pp. 315–326, May 2013.
- [257] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu, “MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features,” *Nucleic Acids Research*, vol. 35, no. Web Server, pp. W339–W344, May 2007.
- [258] L. Breiman, “Random forests,” *Mach Learn*, 2001.
- [259] A. Liaw and M. Wiener, “Classification and regression by randomForest,” *R news*, 2002.
- [260] E. Archer, “rfPermute: Estimate permutation p-values for Random Forest importance metrics. R package version 1.5. 2,” 2013.
- [261] R. Díaz-Uriarte and S. Alvarez de Andrés, “BMC Bioinformatics,” *BMC Bioinformatics*, vol. 7, no. 1, pp. 3–13, 2006.
- [262] R. Díaz-Uriarte and S. A. de Andre, “Variable selection from random forests: application to gene expression data,” *Tech report*, pp. 1–11, Apr. 2017.
- [263] S. Tsuji, Y. Midorikawa, T. Takahashi, K. Yagi, T. Takayama, K. Yoshida, Y. Sugiyama, and H. Aburatani, “Potential responders to FOLFOX therapy for colorectal cancer by Random Forests analysis,” *Br. J. Cancer*, vol. 106, no. 1, pp. 126–132, Nov. 2011.
- [264] P. Nannapaneni, F. Hertwig, M. Depke, M. Hecker, U. Mäder, U. Völker, L. Steil, and S. A. F. T. van Hijum, “Defining the structure of the general stress regulon of *Bacillus subtilis* using targeted microarray analysis and random forest classification.,” *Microbiology (Reading, Engl.)*, vol. 158, no. 3, pp. 696–707, Mar. 2012.
- [265] S. Wuchty, D. Arjona, A. Li, Y. Kotliarov, J. Walling, S. Ahn, A. Zhang, D. Maric, R. Anolik, J. C. Zenklusen, and H. A. Fine, “Prediction of Associations between microRNAs and Gene Expression in Glioma Biology,” *PLoS ONE*, vol. 6, no. 2, pp. e14681–10, Feb. 2011.
- [266] C. S. Bradshaw, A. N. Morton, J. Hocking, S. M. Garland, M. B. Morris, L. M. Moss, L. B. Horvath, I. Kuzevska, and C. K. Fairley, “High recurrence rates of bacterial vaginosis over the course of 12 months after oral metronidazole therapy and factors associated with recurrence.,” *J INFECT DIS*, vol. 193, no. 11, pp. 1478–1486, Jun. 2006.
- [267] J. M. Fettweis, M. G. Serrano, P. H. Girerd, K. K. Jefferson, and G. A. Buck, “A new era of the vaginal microbiome: advances using next-generation sequencing.,” *Chem. Biodivers.*, vol. 9, no. 5, pp. 965–976, May 2012.
- [268] J. M. Marrazzo, D. H. Martin, D. H. Watts, J. Schulte, J. D. Sobel, S. L.

- Hillier, C. Deal, and D. N. Fredricks, "Bacterial vaginosis: identifying research gaps proceedings of a workshop sponsored by DHHS/NIH/NIAID.," presented at the Sexually transmitted diseases, 2010, vol. 37, no. 12, pp. 732–744.
- [269] C. T. Bautista, E. Wurapa, W. B. Sateren, S. Morris, B. Hollingsworth, and J. L. Sanchez, "Bacterial vaginosis: a synthesis of the literature on etiology, prevalence, risk factors, and relationship with chlamydia and gonorrhea infections," *Military Medical Research*, vol. 3, no. 1, pp. 1–10, Feb. 2016.
- [270] S. C. Francis, C. Looker, J. Vandepitte, J. Bukenya, Y. Mayanja, S. Nakubulwa, P. Hughes, R. J. Hayes, H. A. Weiss, and H. Grosskurth, "Bacterial vaginosis among women at high risk for HIV in Uganda: high rate of recurrent diagnosis despite treatment," *Sex Transm Infect*, pp. sextrans–2015–052160–9, Aug. 2015.
- [271] H. C. Wiesenfeld, S. L. Hillier, M. A. Krohn, D. V. Landers, and R. L. Sweet, "Bacterial vaginosis is a strong predictor of *Neisseria gonorrhoeae* and *Chlamydia trachomatis* infection.," *Clin. Infect. Dis.*, vol. 36, no. 5, pp. 663–668, Mar. 2003.
- [272] T. Bertran, P. Brachet, M. Varelle-Delarbre, J. Falenta, A. Dosgilbert, M.-P. Vasson, C. Forestier, A. Tridon, and B. Evrard, "Slight Pro-Inflammatory Immunomodulation Properties of Dendritic Cells by *Gardnerella vaginalis*: The 'Invisible Man' of Bacterial Vaginosis?," *Journal of Immunology Research*, vol. 2016, no. 3125, pp. 1–13, 2016.
- [273] A. R. Thurman, T. Kimble, B. Herold, P. M. M. Mesquita, R. N. Fichorova, H. Y. Dawood, T. Fashemi, N. Chandra, L. Rabe, T. D. Cunningham, S. Anderson, J. Schwartz, and G. Doncel, "Bacterial Vaginosis and Subclinical Markers of Genital Tract Inflammation and Mucosal Immunity," *AIDS Res. Hum. Retroviruses*, p. 150723121229002, Jul. 2015.
- [274] G. Reid, "Cervicovaginal Microbiomes-Threats and Possibilities.," *Trends Endocrinol. Metab.*, Apr. 2016.
- [275] J. R. Marchesi and J. Ravel, "The vocabulary of microbiome research: a proposal," *Microbiome*, vol. 3, no. 1, pp. 1–3, Jul. 2015.
- [276] P. P. Chan and T. M. Lowe, "GtRNAdb: a database of transfer RNA genes detected in genomic sequence.," *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D93–7, Jan. 2009.
- [277] C. J. Yeoman, S. Yildirim, S. M. Thomas, A. S. Durkin, M. Torralba, G. Sutton, C. J. Buhay, Y. Ding, S. P. Dugan-Rocha, D. M. Muzny, X. Qin, R. A. Gibbs, S. R. Leigh, R. Stumpf, B. A. White, S. K. Highlander, K. E. Nelson, and B. A. Wilson, "Comparative genomics of *Gardnerella vaginalis* strains reveals substantial differences in metabolic and virulence potential.," *PLoS ONE*, vol. 5, no. 8, p. e12411, 2010.
- [278] S. E. Reese, K. J. Archer, T. M. Therneau, E. J. Atkinson, C. M. Vachon, M. de Andrade, J. P. A. Kocher, and J. E. Eckel-Passow, "A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis," *Bioinformatics*, vol. 29, no. 22, pp. 2877–2883, Oct. 2013.
- [279] C. M. J. Mentzel, K. Skovgaard, S. Córdoba, J. Herrera Uribe, P. K. Busk,

- and S. Cirera, “Wet-lab tested microRNA assays for qPCR studies with SYBR® Green and DNA primers in pig tissues.,” *Microna*, vol. 3, no. 3, pp. 174–188, 2014.
- [280] S. U. Meyer, M. W. Pfaffl, and S. E. Ulbrich, “Normalization strategies for microRNA profiling experiments: a ‘normal’ way to a hidden layer of complexity?,” *Biotechnol. Lett.*, vol. 32, no. 12, pp. 1777–1788, Dec. 2010.
- [281] K. J. Livak and T. D. Schmittgen, “Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.,” *Methods*, vol. 25, no. 4, pp. 402–408, Dec. 2001.
- [282] E. Archer, *rfPermute: Estimate permutation p-values for Random Forest importance metrics. R package version 1.5. 2.* 2013.
- [283] U. Grömping, “Variable importance assessment in regression: linear regression versus random forest,” *The American Statistician*, 2009.
- [284] C.-H. Chou, N.-W. Chang, S. Shrestha, S.-D. Hsu, Y.-L. Lin, W.-H. Lee, C.-D. Yang, H.-C. Hong, T.-Y. Wei, S.-J. Tu, T.-R. Tsai, S.-Y. Ho, T.-Y. Jian, H.-Y. Wu, P.-R. Chen, N.-C. Lin, H.-T. Huang, T.-L. Yang, C.-Y. Pai, C.-S. Tai, W.-L. Chen, C.-Y. Huang, C.-C. Liu, S.-L. Weng, K.-W. Liao, W.-L. Hsu, and H.-D. Huang, “miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database.,” *Nucleic Acids Research*, vol. 44, no. 1, pp. D239–47, Jan. 2016.
- [285] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.,” *Nat Protoc*, vol. 4, no. 1, pp. 44–57, 2009.
- [286] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, “NIH Image to ImageJ: 25 years of image analysis,” *Nat. Methods*, 2012.
- [287] A. E. Carpenter, T. E. Jones, D. B. Wheeler, and M. Lamprecht, *CellProfiler: open-source, versatile software for high throughput cell image analysis.* Genome Biology, 2006.
- [288] C. Tan, R.-C. Hsia, H. Shou, J. A. Carrasco, R. G. Rank, and P. M. Bavoil, “Variable expression of surface-exposed polymorphic membrane proteins in in vitro-grown Chlamydia trachomatis.,” *Cell Microbiol*, vol. 12, no. 2, pp. 174–187, Feb. 2010.
- [289] L. R. Buckner, D. J. Schust, J. Ding, T. Nagamatsu, W. Beatty, T. L. Chang, S. J. Greene, M. E. Lewis, B. Ruiz, S. L. Holman, R. A. Spagnuolo, R. B. Pyles, and A. J. Quayle, “Innate immune mediator profiles and their regulation in a novel polarized immortalized epithelial cell model derived from human endocervix,” *Journal of Reproductive Immunology*, vol. 92, no. 1, pp. 8–20, Dec. 2011.
- [290] J. E. Allsworth and J. F. Peipert, “Prevalence of Bacterial Vaginosis: 2001–2004 National Health and Nutrition Examination Survey Data,” *Obstet Gynecol*, vol. 109, no. 1, pp. 114–120, Jan. 2007.
- [291] S.-K. Leivonen, A. Rokka, P. Ostling, P. Kohonen, G. L. Corthals, O. Kallioniemi, and M. Perälä, “Identification of miR-193b targets in breast cancer cells and systems biological analysis of their functional impact.,” *Mol. Cell Proteomics*, vol. 10, no. 7, p. M110.005322, Jul. 2011.
- [292] J. Chen, H. E. Feilotter, G. C. Paré, X. Zhang, J. G. W. Pemberton, C.

- Garady, D. Lai, X. Yang, and V. A. Tron, "MicroRNA-193b represses cell proliferation and regulates cyclin D1 in melanoma.," *Am. J. Pathol.*, vol. 176, no. 5, pp. 2520–2529, May 2010.
- [293] S.-K. Leivonen, R. M. A. K. auml, P. O. stling, P. Kohonen, S. Haapa-Paananen, K. Kleivi, E. Enerly, A. Aakula, K. H. O. m, N. Sahlberg, V. N. Kristensen, A.-L. B. O. rresen-Dale, P. Saviranta, M. P. A. L. auml, and O. Kallioniemi, "Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines," *Oncogene*, vol. 28, no. 44, pp. 3926–3936, Aug. 2009.
- [294] C. Xu, S. Liu, H. Fu, S. Li, Y. Tie, J. Zhu, R. Xing, Y. Jin, Z. Sun, and X. Zheng, "MicroRNA-193b regulates proliferation, migration and invasion in human hepatocellular carcinoma cells.," *Eur. J. Cancer*, vol. 46, no. 15, pp. 2828–2836, Oct. 2010.
- [295] C. Gastaldi, T. Bertero, N. Xu, I. Bourget-Ponzio, K. Lebrigand, S. Fourre, A. Popa, N. Cardot-Leccia, G. Meneguzzi, E. Sonkoly, A. Pivarcsi, B. Mari, P. Barbry, G. Ponzio, and R. Rezzonico, "miR-193b/365a cluster controls progression of epidermal squamous cell carcinoma.," *Carcinogenesis*, vol. 35, no. 5, pp. 1110–1120, May 2014.
- [296] M. Lenarduzzi, A. B. Y. Hui, N. M. Alajez, W. Shi, J. Williams, S. Yue, B. O'Sullivan, and F.-F. Liu, "MicroRNA-193b enhances tumor progression via down regulation of neurofibromin 1.," *PLoS ONE*, vol. 8, no. 1, p. e53765, 2013.
- [297] X.-F. Li, P.-J. Yan, and Z.-M. Shao, "Downregulation of miR-193b contributes to enhance urokinase-type plasminogen activator (uPA) expression and tumor progression and invasion in human breast cancer.," *Oncogene*, vol. 28, no. 44, pp. 3937–3948, Nov. 2009.
- [298] H. Zhou, K. Wang, Z. Hu, and J. Wen, "TGF-beta1 alters microRNA profile in human gastric cancer cells.," *Chin J Cancer Res*, vol. 25, no. 1, pp. 102–111, Feb. 2013.
- [299] Q. Zhong, T. Wang, P. Lu, R. Zhang, J. Zou, and S. Yuan, "miR-193b promotes cell proliferation by targeting Smad3 in human glioma.," *Journal of Neuroscience Research*, vol. 92, no. 5, pp. 619–626, Feb. 2014.
- [300] G. Zhu, S. E. Conner, X. Zhou, C. Shih, T. Li, B. D. Anderson, H. B. Brooks, R. M. Campbell, E. Considine, J. A. Dempsey, M. M. Faul, C. Ogg, B. Patel, R. M. Schultz, C. D. Spencer, B. Teicher, and S. A. Watkins, "Synthesis, structure-activity relationship, and biological studies of indolocarbazoles as potent cyclin D1-CDK4 inhibitors.," *J. Med. Chem.*, vol. 46, no. 11, pp. 2027–2030, May 2003.
- [301] R. Soni, L. Muller, P. Furet, J. Schoepfer, C. Stephan, S. Zumstein-Mecker, H. Fretz, and B. Chaudhuri, "Inhibition of cyclin-dependent kinase 4 (Cdk4) by fascaplysin, a marine natural product.," *Biochem. Biophys. Res. Commun.*, vol. 275, no. 3, pp. 877–884, Sep. 2000.
- [302] A. S. Zevin, I. Y. Xie, K. Birse, K. Arnold, L. Romas, G. Westmacott, R. M. Novak, S. McCorrister, L. R. McKinnon, C. R. Cohen, R. Mackelprang, J. Lingappa, D. A. Lauffenburger, N. R. Klatt, and A. D. Burgener, "Microbiome Composition and Function Drives Wound-Healing

- Impairment in the Female Genital Tract,” *PLoS Pathog*, vol. 12, no. 9, pp. e1005889–20, Sep. 2016.
- [303] R. N. Fichorova and J. G. Rheinwald, *Generation of papillomavirus-immortalized cell lines from normal human ectocervical, endocervical, and vaginal epithelium that maintain expression of ...* *Biology of ...*, 1997.
- [304] M. Bens, A. Bogdanova, F. Cluzeaud, L. Miquerol, S. Kerneis, J. P. Kraehenbuhl, A. Kahn, E. Pringault, and A. Vandewalle, “Transimmortalized mouse intestinal cells (m-ICc12) that maintain a crypt phenotype,” *American Journal of Physiology - Cell Physiology*, vol. 270, no. 6, pp. C1666–C1674, Jun. 1996.
- [305] N. M. Gilbert, W. G. Lewis, and A. L. Lewis, “Clinical Features of Bacterial Vaginosis in a Murine Model of Vaginal Infection with *Gardnerella vaginalis*,” *PLoS ONE*, vol. 8, no. 3, pp. e59539–13, Mar. 2013.
- [306] J. Wang, “Bacterial vaginosis,” *Primary Care Update for OB/GYNS*, vol. 7, no. 5, pp. 181–185, Sep. 2000.
- [307] J. L. Patterson, A. Stull-Lane, P. H. Girerd, and K. K. Jefferson, “Analysis of adherence, biofilm formation and cytotoxicity suggests a greater virulence potential of *Gardnerella vaginalis* relative to other bacterial-vaginosis-associated anaerobes.,” *Microbiology (Reading, Engl.)*, vol. 156, no. 2, pp. 392–399, Feb. 2010.
- [308] K. Nishiyama, M. Sugiyama, and T. Mukai, “Adhesion Properties of Lactic Acid Bacteria on Intestinal Mucin,” *Microorganisms*, vol. 4, no. 4, pp. 34–18, Dec. 2016.
- [309] M. Kim, H. Ashida, M. Ogawa, Y. Yoshikawa, H. Mimuro, and C. Sasakawa, “Bacterial interactions with the host epithelium.,” *Cell Host Microbe*, vol. 8, no. 1, pp. 20–35, Jul. 2010.
- [310] S. Boris, J. E. Suárez, F. Vázquez, and C. Barbés, “Adherence of human vaginal lactobacilli to vaginal epithelial cells and interaction with uropathogens.,” *Infection and Immunity*, vol. 66, no. 5, pp. 1985–1989, May 1998.
- [311] S. S. Witkin, H. Mendes-Soares, I. M. Linhares, A. Jayaram, W. J. Ledger, and L. J. Forney, “Influence of vaginal bacteria and D- and L-lactic acid isomers on vaginal extracellular matrix metalloproteinase inducer: implications for protection against upper genital tract infections.,” *mBio*, vol. 4, no. 4, pp. e00460–13–e00460–13, 2013.
- [312] S. Miyagawa and T. Iguchi, “Epithelial estrogen receptor 1 intrinsically mediates squamous differentiation in the mouse vagina.,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 42, pp. 12986–12991, Oct. 2015.
- [313] H. E. Averette, G. D. Weinstein, and P. Frost, “Autoradiographic analysis of cell proliferation kinetics in human genital tissues. I. Normal cervix and vagina.,” *YMOB*, vol. 108, no. 1, pp. 8–17, Sep. 1970.
- [314] M. Xia, R. E. Bumgarner, M. F. Lampe, and W. E. Stamm, “Chlamydia trachomatis infection alters host cell transcription in diverse cellular pathways.,” *JINFECT DIS*, vol. 187, no. 3, pp. 424–434, Feb. 2003.

- [315] C. Elwell, K. Mirrashidi, and J. Engel, “Chlamydia cell biology and pathogenesis,” *Nature Publishing Group*, vol. 14, no. 6, pp. 385–400, Apr. 2016.
- [316] A. L. Patel, X. Chen, S. T. Wood, E. S. Stuart, K. F. Arcaro, D. P. Molina, S. Petrovic, C. M. Furdui, and A. W. Tsang, “Activation of epidermal growth factor receptor is required for Chlamydia trachomatis development,” *BMC Microbiol.*, vol. 14, no. 1, p. 277, Dec. 2014.
- [317] C. Chumduri, R. K. Gurumurthy, P. K. Zadora, Y. Mi, and T. F. Meyer, “Chlamydia Infection Promotes Host DNA Damage and Proliferation but Impairs the DNA Damage Response,” *Cell Host Microbe*, vol. 13, no. 6, pp. 746–758, Jun. 2013.
- [318] B. P. Katz, “Estimating transmission probabilities for chlamydial infection,” *Statist. Med.*, vol. 11, no. 5, pp. 565–577, Mar. 1992.
- [319] E. Lycke, G. B. Löwhagen, G. Hallhagen, G. Johannisson, and K. Ramstedt, “The risk of transmission of genital Chlamydia trachomatis infection is less than that of genital Neisseria gonorrhoeae infection,” *Sexually Transmitted Diseases*, vol. 7, no. 1, pp. 6–10, Jan. 1980.
- [320] T. C. Quinn, C. Gaydos, M. Shepherd, L. Bobo, E. W. Hook, R. Viscidi, and A. Rompalo, “Epidemiologic and microbiologic correlates of Chlamydia trachomatis infection in sexual partnerships,” *JAMA*, vol. 276, no. 21, pp. 1737–1742, Dec. 1996.
- [321] F. Y. S. Kong and J. S. Hocking, “Treatment challenges for urogenital and anorectal Chlamydia trachomatis,” *BMC Infect. Dis.*, vol. 15, no. 1, p. 293, Jul. 2015.
- [322] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: A flexible trimmer for Illumina Sequence Data.”
- [323] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions,” *Genome biology*, vol. 14, no. 4, p. R36, Apr. 2013.
- [324] NCBI Resource Coordinators, “Database Resources of the National Center for Biotechnology Information,” *Nucleic Acids Research*, vol. 45, no. 1, pp. D12–D17, Jan. 2017.
- [325] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the royal statistical society Series B (...)*, 1995.
- [326] J. J. Goeman and A. Solari, “Multiple hypothesis testing in genomics,” *Statist. Med.*, vol. 33, no. 11, pp. 1946–1978, Jan. 2014.
- [327] A. Krämer, J. Green, J. Pollard, and S. Tugendreich, “Causal analysis approaches in Ingenuity Pathway Analysis,” *Bioinformatics*, vol. 30, no. 4, pp. 523–530, Feb. 2014.
- [328] S. THIAGALINGAM, K. H. CHENG, H. J. LEE, N. MINEVA, A. THIAGALINGAM, and J. F. PONTE, “Histone Deacetylases: Unique Players in Shaping the Epigenetic Histone Code,” *Annals of the New York Academy of Sciences*, vol. 983, no. 1, pp. 84–100, Mar. 2003.
- [329] D. Mottet, S. Pirotte, V. Lamour, M. Hagedorn, S. Javerzat, A. Bikfalvi, A.

- Bellaïche, E. Verdin, and V. Castronovo, "HDAC4 represses p21WAF1/Cip1 expression in human cancer cells through a Sp1-dependent, p53-independent mechanism," *Oncogene*, vol. 28, no. 2, pp. 243–256, Oct. 2008.
- [330] M. A. Hamon and P. Cossart, "Histone modifications and chromatin remodeling during bacterial infections.," *Cell Host Microbe*, vol. 4, no. 2, pp. 100–109, Aug. 2008.
- [331] E. E. Hull, M. R. Montgomery, and K. J. Leyva, "HDAC Inhibitors as Epigenetic Regulators of the Immune System: Impacts on Cancer Therapy and Inflammatory Diseases.," *BioMed Research International*, vol. 2016, no. 4, pp. 8797206–15, 2016.
- [332] R. C. E. a-Oliveira, J. E. L. I. S. Fachi, A. Vieira, F. T. Sato, and M. A. E. L. R. Vinolo, "Regulation of immune cell function by short-chain fatty acids," *Clin Trans Immunol*, vol. 5, no. 4, pp. e73–8, Apr. 2016.
- [333] P. V. Chang, L. Hao, S. Offermanns, and R. Medzhitov, "The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition.," *Proceedings of the National Academy of Sciences*, vol. 111, no. 6, pp. 2247–2252, Feb. 2014.
- [334] P. Mirmonsef, M. R. Zariffard, D. Gilbert, H. Makinde, A. L. Landay, and G. T. Spear, "Short-Chain Fatty Acids Induce Pro-Inflammatory Cytokine Production Alone and in Combination with Toll-Like Receptor Ligands," *American Journal of Reproductive Immunology*, vol. 67, no. 5, pp. 391–400, Nov. 2011.
- [335] S. Tedelind, F. Westberg, M. Kjerrulf, and A. Vidal, *Anti-inflammatory properties of the short-chain fatty acids acetate and propionate: a study with relevance to inflammatory bowel disease*. World Journal of Gastroenterology, 2007.
- [336] N. Gupta, P. M. Martin, P. D. Prasad, and V. Ganapathy, "SLC5A8 (SMCT1)-mediated transport of butyrate forms the basis for the tumor suppressive function of the transporter," *Life Sciences*, vol. 78, no. 21, pp. 2419–2425, Apr. 2006.
- [337] A. Borthakur, A. N. Anbazhagan, A. Kumar, G. Raheja, V. Singh, K. Ramaswamy, and P. K. Dudeja, "The probiotic *Lactobacillus plantarum* counteracts TNF- α -induced downregulation of SMCT1 expression and function.," *Am. J. Physiol. Gastrointest. Liver Physiol.*, vol. 299, no. 4, pp. G928–34, Oct. 2010.
- [338] A. P. HALESTRAP and N. T. PRICE, "The proton-linked monocarboxylate transporter (MCT) family: structure, function and regulation," *Biochemical Journal*, vol. 343, no. 2, pp. 281–299, Oct. 1999.
- [339] R. K. Gill, S. Saksena, W. A. Alrefai, Z. Sarwar, J. L. Goldstein, R. E. Carroll, K. Ramaswamy, and P. K. Dudeja, "Expression and membrane localization of MCT isoforms along the length of the human intestine.," *American Journal of Physiology - Cell Physiology*, vol. 289, no. 4, pp. C846–52, Oct. 2005.
- [340] M. A. Glozak and E. Seto, "Histone deacetylases and cancer.," *Oncogene*, vol. 26, no. 37, pp. 5420–5432, Aug. 2007.

- [341] W.-I. Choi, B.-N. Jeon, J.-H. Yoon, D.-I. Koh, M.-H. Kim, M.-Y. Yu, K.-M. Lee, Y. Kim, K. Kim, S. S. Hur, C.-E. Lee, K.-S. Kim, and M.-W. Hur, "The proto-oncoprotein FBI-1 interacts with MBD3 to recruit the Mi-2/NuRD-HDAC complex and BCoR and to silence p21WAF/CDKN1A by DNA methylation.," *Nucleic Acids Research*, vol. 41, no. 13, pp. 6403–6420, Jul. 2013.
- [342] C. Wang, M. Fu, R. H. Angeletti, L. Siconolfi-Baez, A. T. Reutens, C. Albanese, M. P. Lisanti, B. S. Katzenellenbogen, S. Kato, T. Hopp, S. A. Fuqua, G. N. Lopez, P. J. Kushner, and R. G. Pestell, "Direct acetylation of the estrogen receptor alpha hinge region by p300 regulates transactivation and hormone sensitivity.," *Journal of Biological Chemistry*, vol. 276, no. 21, pp. 18375–18383, May 2001.
- [343] D. W. Hilbert, W. L. Smith, S. G. Chadwick, G. Toner, E. Mordechai, M. E. Adelson, J. D. Sobel, and S. E. Gyax, "Development and validation of a highly accurate quantitative real-time PCR assay for the diagnosis of bacterial vaginosis.," *J. Clin. Microbiol.*, pp. JCM.03104–15, Jan. 2016.
- [344] A. L. Muhleisen and M. M. Herbst-Kralovetz, "Menopause and the vaginal microbiome," *Maturitas*, vol. 91, pp. 42–50, Sep. 2016.
- [345] E. R. Shamir and A. J. Ewald, "Three-dimensional organotypic culture: experimental models of mammalian biology and disease.," *Nat. Rev. Mol. Cell Biol.*, vol. 15, no. 10, pp. 647–664, Oct. 2014.
- [346] H. J. Kim, D. Huh, G. Hamilton, and D. E. Ingber, "Human gut-on-a-chip inhabited by microbial flora that experiences intestinal peristalsis-like motions and flow," *Lab Chip*, vol. 12, no. 12, pp. 2165–10, 2012.
- [347] C. S. Velu and H. L. Grimes, "Utilizing AntagomiR (Antisense microRNA) to Knock Down microRNA in Murine Bone Marrow Cells," in *Rational Drug Design*, no. 15, Totowa, NJ: Humana Press, 2012, pp. 185–195.
- [348] T. K. Kim and J. H. Eberwine, "Mammalian cell transfection: the present and the future," *Anal Bioanal Chem*, vol. 397, no. 8, pp. 3173–3178, Jun. 2010.
- [349] B. L. Kinlock, Y. Wang, T. M. Turner, C. Wang, and B. Liu, "Transcytosis of HIV-1 through Vaginal Epithelial Cells Is Dependent on Trafficking to the Endocytic Recycling Pathway," *PLoS ONE*, vol. 9, no. 5, pp. e96760–11, May 2014.
- [350] E. Taneva, K. Crooker, S. H. Park, J. T. Su, A. Ott, N. Cheshenko, I. Szleifer, P. F. Kiser, B. Frank, P. M. M. Mesquita, and B. C. Herold, "Differential Mechanisms of Tenofovir and Tenofovir Disoproxil Fumarate Cellular Transport and Implications for Topical Preexposure Prophylaxis," *Antimicrob. Agents Chemother.*, vol. 60, no. 3, pp. 1667–1675, Feb. 2016.
- [351] T. Clément, V. Salone, and M. Rederstorff, "Dual Luciferase Gene Reporter Assays to Study miRNA Function," in *Small Non-Coding RNAs*, vol. 1296, no. 17, M. Rederstorff, Ed. New York, NY: Springer New York, 2015, pp. 187–198.
- [352] H. J. Kim and S.-C. Bae, "Histone deacetylase inhibitors: molecular mechanisms of action and clinical trials as anti-cancer drugs.," *Am J Transl Res*, vol. 3, no. 2, pp. 166–179, Feb. 2011.

- [353] W. S. Xu, R. B. Parmigiani, and P. A. Marks, "Histone deacetylase inhibitors: molecular mechanisms of action.," *Oncogene*, vol. 26, no. 37, pp. 5541–5552, Aug. 2007.