Vaginal microbiota dynamics and pubertal development

Roxana J. Hickey

Contents

Description

This is a supplement to the paper "Vaginal microbiota of adolescent girls resemble those of reproductive-age women prior to the onset of menarche" by Hickey et al. Please refer to the paper for complete information about the objectives and study design.

The embedded R code works through the first set of analyses and generation of figures related to the assessment of vaginal microbiota composition in girls and mothers. The analyses can be run directly from the R Markdown file using RStudio. It should be run after "01-data-prep.Rmd" and "02-hclust-pcoa.Rmd".

Objective

Earlier analyses dealt with characterizing vaginal microbiota composition and identifying major groups using clustering and ordination approaches. Next, we will take a closer look at the dynamics of these communities over time as girls progress through puberty and menarche. Special attention is devoted to assessing trends in the relative abundance of lactic acid bacteria and vaginal pH, as these are generally considered indicators of a 'healthy' vaginal microbiota in reproductive age women.

Initial setup

Clear the workspace, load data from the previous step, and load necessary packages.

Note: If you run the R Markdown script 'as is' from the same directory containing it and the 'data' and 'data-postproc' subdirectories, all figures will be printed inside the resulting PDF or HTML output. If you want to save the figures as individual files, uncomment the lines below starting with 'dir.create()' as well as any lines throughout the script starting with 'ggsave()' or 'pdf()'. I made note of each of these within the chunk code.

```
## Clear current workspace
rm(list=ls())

## Load the RData file created from adolescent-supp-02.Rmd
## (this contains data from 01-data-prep as well)
load("data-postproc/02-hclust-pcoa-last-run.RData")

## Load packages
library(ape)
```

```
## Warning: package 'ape' was built under R version 3.1.1
```

```
library(ggplot2)
library(gplots)
```

```
## Warning: package 'gplots' was built under R version 3.1.1

## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009
##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(grid)
library(gridExtra)
library(reshape)
library(scales)
```

```
## Function to get rid of pesky auto-factor (acknowledgement: Matthew W. Pennell, University of Idaho)
unfactor <- function(x){
  as.character(levels(x))[x]
}
```

I: Summarize community dynamics of vaginal (and vulvar) microbiota

First we will create an individual summary for each individual participant (and her mother, if applicable) to look at community composition of the vaginal and vulvar microbiota, along with selected metadata, at every visit during the study. The output is a PDF file with each page containing the following plots for a given subject:

- Community composition of girl's vaginal microbiota
- Community composition of girl's vulvar microbiota
- Community composition of mother's vaginal microbiota
- Girl's menarcheal status
- Girl's Tanner breast and pubic stages (clinician-assessed)
- Vaginal pH of girl and mother (when available)

```
## g_legend function from http://stackoverflow.com/questions/11883844/
## inserting-a-table-under-the-legend-in-a-ggplot2-histogram
g_legend <- function(a.gplot){
  tmp <- ggplot_gtable(ggplot_build(a.gplot))
  leg <- which(sapply(tmp$grobs, function(x) x$name) == "guide-box")
  legend <- tmp$grobs[[leg]]
  return(legend)}

# source("scripts/community_dynamics_profiles.R")
```

Highlight examples of *Lactobacillus* dominant vaginal microbiota

Next we will plot just a few examples of vaginal microbiota with *Lactobacillus* predominant by menarche. I have selected subjects 102, 103, 107 and 109. We will plot only the vaginal microbiota composition, menarche status, and Tanner breast stage for each. This also requires a bit of setup:

```
## Subset prop and meta for subjects 102, 103, 107 and 109
meta.fig5 <- subset(meta, subject %in% c(102, 103, 107, 109) & site=="vag")
prop.fig5 <- prop.red[,colnames(prop.red) %in% rownames(meta.fig5)]

taxa.pick40 <- order(rowMeans(as.matrix(prop.fig5), na.rm=TRUE),
                     decreasing=T)[1:40]
prop.fig5.top40 <- prop.fig5[taxa.pick40,]

mp.fig5 <- data.frame(cbind(as.character(meta.fig5$subject),
                            meta.fig5$visit,
                            as.character(meta.fig5$men.stat),
```

```
                            meta.fig5$tan.br.dr,
                            t(prop.fig5.top40)))
colnames(mp.fig5)[1:4] <- c("subject",
                            "visit",
                            "men.stat",
                            "tan.br.dr")

mp.fig5.lg <- melt(mp.fig5, id.vars=c("subject",
                                      "visit",
                                      "men.stat",
                                      "tan.br.dr"))

mp.fig5.lg$visit <- as.integer(unfactor(mp.fig5.lg$visit))
mp.fig5.lg$value <- as.numeric(unfactor(mp.fig5.lg$value))

mp.fig5.lg$tan.br.dr <- factor(mp.fig5.lg$tan.br.dr, levels=c(1,2,3,4,5))

mp.s102 <- subset(mp.fig5.lg, subject==102)
mp.s103 <- subset(mp.fig5.lg, subject==103)
mp.s107 <- subset(mp.fig5.lg, subject==107)
mp.s109 <- subset(mp.fig5.lg, subject==109)

## Plot vaginal microbiota composition of subject 102 (w/ legend)
gg.taxa.bar.102 <- ggplot(mp.s102, aes(x=visit, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  scale_fill_manual(values=col.taxa[taxa.pick40],
                    name="Taxon") +
  scale_x_discrete(limits=seq(1, max(mp.s102$visit), by=1),
                   name="") +
  ylab("Proportion") +
  ggtitle("Subject 102") +
  theme_cust +
  guides(fill=guide_legend(ncol=8)) +
  theme(axis.text.x=element_blank(),
        legend.text=element_text(size=7),
        legend.title=element_text(size=8),
        legend.position=c(0,0),
        legend.justification=c(0,0),
        legend.direction="horizontal")

## Plot vaginal microbiota composition of subject 103
gg.taxa.bar.103 <- ggplot(mp.s103, aes(x=visit, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  scale_fill_manual(values=col.taxa[taxa.pick40], name="Taxon") +
  scale_x_discrete(limits=seq(1, max(mp.s103$visit), by=1),
                   name="") +
  ylab("Proportion") +
  ggtitle("Subject 103") +
  theme_cust +
  theme(axis.text.x=element_blank(),
        legend.position="none")

## Plot vaginal microbiota composition of subject 107
```

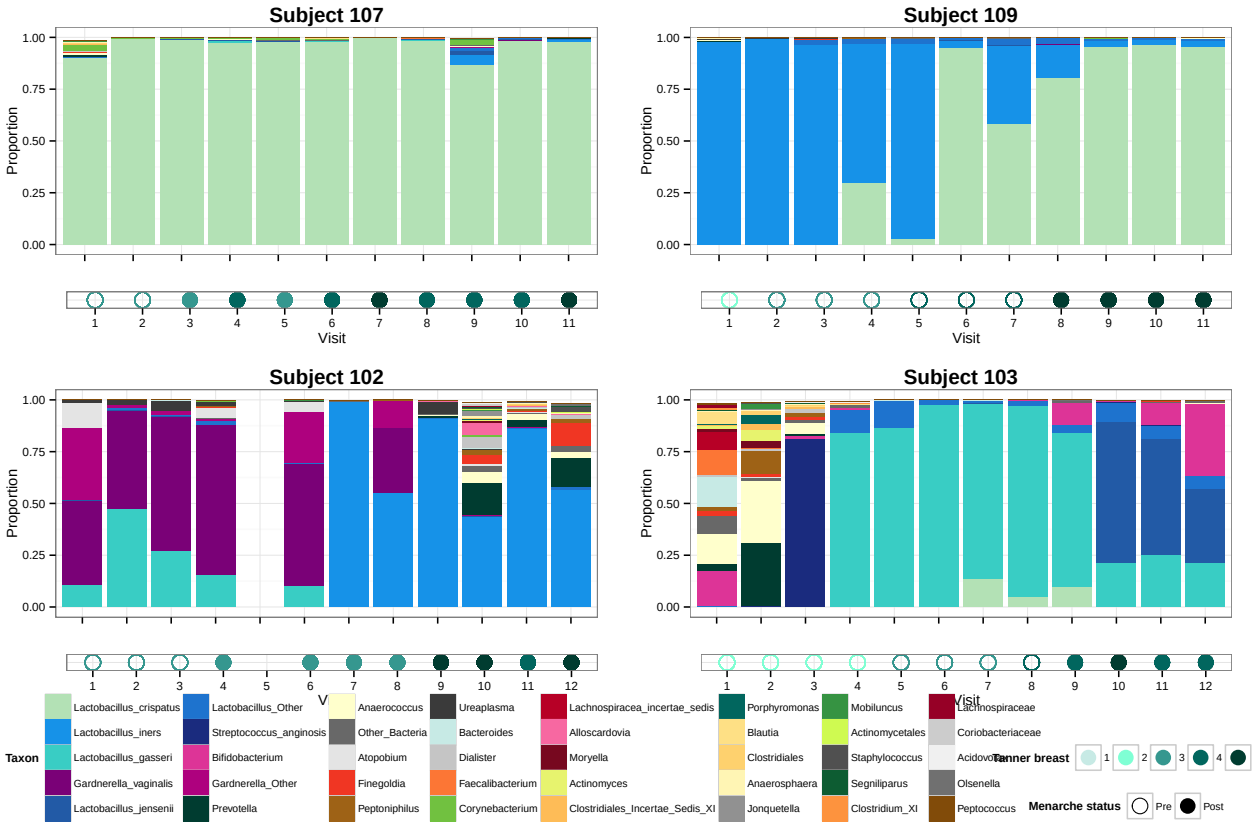
```
gg.taxa.bar.107 <- ggplot(mp.s107, aes(x=visit, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  scale_fill_manual(values=col.taxa[taxa.pick40], name="Taxon") +
  scale_x_discrete(limits=seq(1, max(mp.s107$visit), by=1),
                   name="") +
  ylab("Proportion") +
  ggtitle("Subject 107") +
  theme_cust +
  theme(axis.text.x=element_blank(),
       legend.position="none")

## Plot vaginal microbiota composition of subject 109
gg.taxa.bar.109 <- ggplot(mp.s109, aes(x=visit, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  scale_fill_manual(values=col.taxa[taxa.pick40], name="Taxon") +
  scale_x_discrete(limits=seq(1, max(mp.s109$visit), by=1),
                   name="") +
  ylab("Proportion") +
  ggtitle("Subject 109") +
  theme_cust +
  theme(axis.text.x=element_blank(),
        legend.position="none")

## Plot Tanner/menarche of subject 102 (w/ legend)
gg.tb.dot.102 <- ggplot(mp.s102, aes(x=visit, y=subject,
                                     shape=men.stat, color=tan.br.dr)) +
  geom_point(size=5) +
  scale_color_manual(values=col.tanner,
                     name="Tanner breast",
                     drop=FALSE) +
  scale_shape_manual(breaks=c("pre", "post"),
                     labels=c("Pre", "Post"),
                     values=c(16,1),
                     name="Menarche status") +
  scale_x_discrete(limits=seq(1, max(mp.s102$visit), by=1),
                   name="Visit") +
  ylab("") +
  theme_cust +
  theme(axis.ticks.y=element_blank(),
        axis.text.y=element_blank(),
        legend.text=element_text(size=7),
        legend.title=element_text(size=8),
        legend.position=c(0.8,0.5),
        legend.justification=c(0.5,0.5),
        legend.direction="horizontal")

## Plot Tanner/menarche of subject 103
gg.tb.dot.103 <- ggplot(mp.s103, aes(x=visit, y=subject,
                                     shape=men.stat, color=tan.br.dr)) +
  geom_point(size=5) +
  scale_color_manual(values=col.tanner,
                     name="Tanner breast",
                     drop=FALSE) +
```

```
  scale_shape_manual(breaks=c("pre", "post"),
                     labels=c("Pre", "Post"),
                     values=c(16,1),
                     name="Menarche status") +
  scale_x_discrete(limits=seq(1, max(mp.s103$visit), by=1),
                   name="Visit") +
  ylab("") +
  theme_cust +
  theme(axis.ticks.y=element_blank(),
        axis.text.y=element_blank(),
        legend.position="none")

## Plot Tanner/menarche of subject 107
gg.tb.dot.107 <- ggplot(mp.s107, aes(x=visit, y=subject,
                                     shape=men.stat, color=tan.br.dr)) +
  geom_point(size=5) +
  scale_color_manual(values=col.tanner,
                     name="Tanner breast",
                     drop=FALSE) +
  scale_shape_manual(breaks=c("pre", "post"),
                     labels=c("Pre", "Post"),
                     values=c(16,1),
                     name="Menarche status") +
  scale_x_discrete(limits=seq(1, max(mp.s107$visit), by=1),
                   name="Visit") +
  ylab("") +
  theme_cust +
  theme(axis.ticks.y=element_blank(),
        axis.text.y=element_blank(),
        legend.position="none")

## Plot Tanner/menarche of subject 109
gg.tb.dot.109 <- ggplot(mp.s109, aes(x=visit, y=subject,
                                     shape=men.stat, color=tan.br.dr)) +
  geom_point(size=5) +
  scale_color_manual(values=col.tanner,
                     name="Tanner breast",
                     drop=FALSE) +
  scale_shape_manual(breaks=c("pre", "post"),
                     labels=c("Pre", "Post"),
                     values=c(16,1),
                     name="Menarche status") +
  scale_x_discrete(limits=seq(1, max(mp.s109$visit), by=1),
                   name="Visit") +
  ylab("") +
  theme_cust +
  theme(axis.ticks.y=element_blank(),
        axis.text.y=element_blank(),
        legend.position="none")
```

Figure 5. Transitions to *Lactobacillus*-dominant vaginal microbiota in four peri-menarcheal girls.

Each panel shows the vaginal bacterial community profiles and associated pubertal development of four participants sampled longitudinally. The top bar plot of each panel shows the proportions of bacterial taxa present at each sampling, with colors of taxa indicated in the legend at bottom left. Below each bar plot the menarcheal status (M) and Tanner breast (TB) scores as assessed by a clinician are indicated following the color scheme shown in the legend at bottom right. Empty spaces in the plots indicate a skipped scheduled quarterly visit or a visit in which a vaginal swab or metadata were not collected.

```
## Uncomment next line to save as a PDF
# pdf("figures/fig-5-community-dynamics.pdf", width=12, height=8, pointsize=8, bg="transparent")
grid.arrange(gg.taxa.bar.107 + theme(plot.margin=unit(c(0.25, 1, 0, 0.8), "lines")),
             gg.taxa.bar.109 + theme(plot.margin=unit(c(0.25, 1, 0, 0.8), "lines")),
             gg.tb.dot.107 + theme(plot.margin=unit(c(0.25, 1, 0.5, 2.4), "lines")),
             gg.tb.dot.109 + theme(plot.margin=unit(c(0.25, 1, 0.5, 2.4), "lines")),
             gg.taxa.bar.102 + theme(legend.position="none",
                                     plot.margin=unit(c(0.25, 1, 0, 0.8), "lines")),
             gg.taxa.bar.103 + theme(plot.margin=unit(c(0.25, 1, 0, 0.8), "lines")),
             gg.tb.dot.102 + theme(legend.position="none",
                                   plot.margin=unit(c(0.25, 1, 0.5, 2.4), "lines")),
             gg.tb.dot.103 + theme(plot.margin=unit(c(0.25, 1, 0.5, 2.4), "lines")),
             g_legend(gg.taxa.bar.102),
             g_legend(gg.tb.dot.102),
             ncol=2, heights=c(3,0.75,3,0.75,1.25))
```



```
dev.off()
```

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

II. *Gardnerella* in perimenarcheal vaginal microbiota

Gardnerella vaginalis was surprisingly common among girls in our study. This is notable because the species (or genus) is commonly associated with bacterial vaginosis, and some have argued it is acquired through sexual contact. However, the girls in this study had no history of sexual activity and reported themselves in good health. Below we plot the changes in proportion of *Gardnerella* over time in girls who had at least 5% *Gardnerella* in her vaginal microbiota at any point in time during the study.

```
meta.vag$Gvag <- spe.prop.vag[,"Gardnerella_vaginalis"]
meta.vag$Gardnerella <- rowSums(spe.prop.vag[,grep("Gardnerella", colnames(spe.prop.vag))])

meta.vag.gard.10 <- subset(meta.vag, type=="girl" & Gardnerella >= 0.05)
unique(meta.vag.gard.10$subject) # n=11
```

```
##  [1] 102 105 112 115 116 120 121 124 126 129 135
## 55 Levels: 101 102 103 104 105 106 107 108 109 110 111 112 113 114 ... 235
```

```
## We want to plot Gardnerella over time in all of the girls who carried it at some point,
## so we need to grab all the obervations for the girls listed above.
meta.vag.gard.10.fullobs <- meta.vag[meta.vag$subject %in% meta.vag.gard.10$subject,]
```

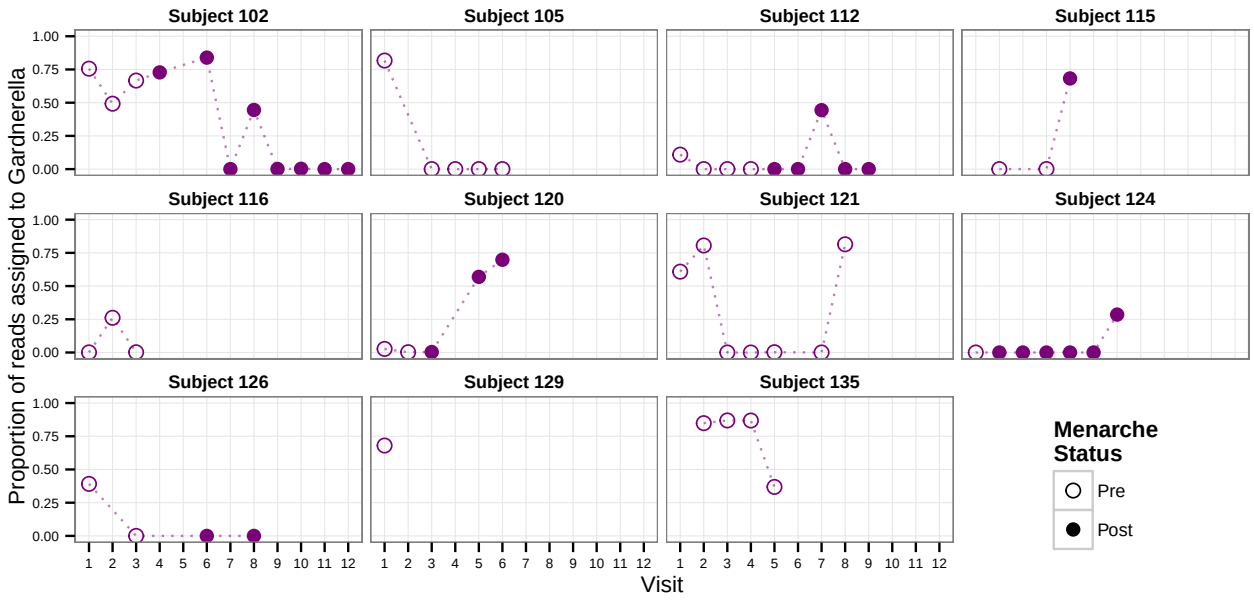
Figure S3. Proportion of *Gardnerella* over time in the vaginal microbiota of girls.

Gardnerella was present in the vaginal microbiota at a proportion of 0.10 or greater at least once in 11/31 adolescent participants. Each subplot shows the proportion of *Gardnerella* (encompassing sequence reads assigned to either the species level as *G. vaginalis* or genus level as *Gardnerella*) in the vaginal microbiota of a single participant at each clinical visit. Open circles represent premenarcheal samples, and filled circles represent postmenarcheal samples. The x-axis indicates the clinical visit at which each sample was collected; visits occurred approximately every three months.

```
gg.girl.gard <- ggplot(meta.vag.gard.10.fullobs,
                       aes(x=visit, y=Gardnerella, group=subject,
                           color=type, shape=men.stat))
gg.girl.gard + geom_line(lty=3, alpha=0.5) +
  facet_wrap( ~ gm.pair.fullID, ncol=4) +
  geom_point(size=3) +
  scale_color_manual(values="#7A0177", guide=FALSE) +
  scale_shape_manual(values=c(16, 1),
                     breaks=c("pre", "post"),
                     labels=c("Pre", "Post"),
                     name="Menarche\nStatus", na.value=10) +
  xlab("Visit") +
  ylab("Proportion of reads assigned to Gardnerella") +
```

```
  scale_x_discrete(1:14, name="Visit") +
  ylim(0,1) +
  theme_cust_nominor +
  theme(axis.text=element_text(size=6),
        legend.justification=c(1,1),
        legend.position=c(0.95,0.3))
```

```
## geom_path: Each group consist of only one observation. Do you need to adjust the group aesthetic?
```




```
## Uncomment next line to save as a PDF
# ggsave("supplemental/fig-s3-gardnerella-vag.pdf", width=8, height=4, units="in")
```

Gardnerella in vaginal and vulvar microbiota

I was also curious to see whether the same patterns were observed in the vulva. Not surprisingly, they are fairly similar, although in some cases the vulva contained a much higher proportion of *Gardnerella* than the vagina.

```
meta$Gvag <- t(prop.red["Gardnerella_vaginalis",])
meta$Gardnerella <- colSums(prop.red[grep("Gardnerella", rownames(prop.red)),])

meta.girl.gard.10 <- subset(meta, type=="girl" & Gardnerella >= 0.10)
unique(meta.girl.gard.10$subject) # n=11
```

```
##  [1] 102 105 112 115 116 120 121 124 126 129 135
## 55 Levels: 101 102 103 104 105 106 107 108 109 110 111 112 113 114 ... 235
```

```
## Subjects who had Gardnerella in vagina
unique(meta.girl.gard.10$subject[meta.girl.gard.10$site=="vag"]) # n=11
```

```
##  [1] 102 105 112 115 116 120 121 124 126 129 135
## 55 Levels: 101 102 103 104 105 106 107 108 109 110 111 112 113 114 ... 235
```

```
## Subjects who had Gardnerella on vulva
unique(meta.girl.gard.10$subject[meta.girl.gard.10$site=="vul"]) # n=9
```

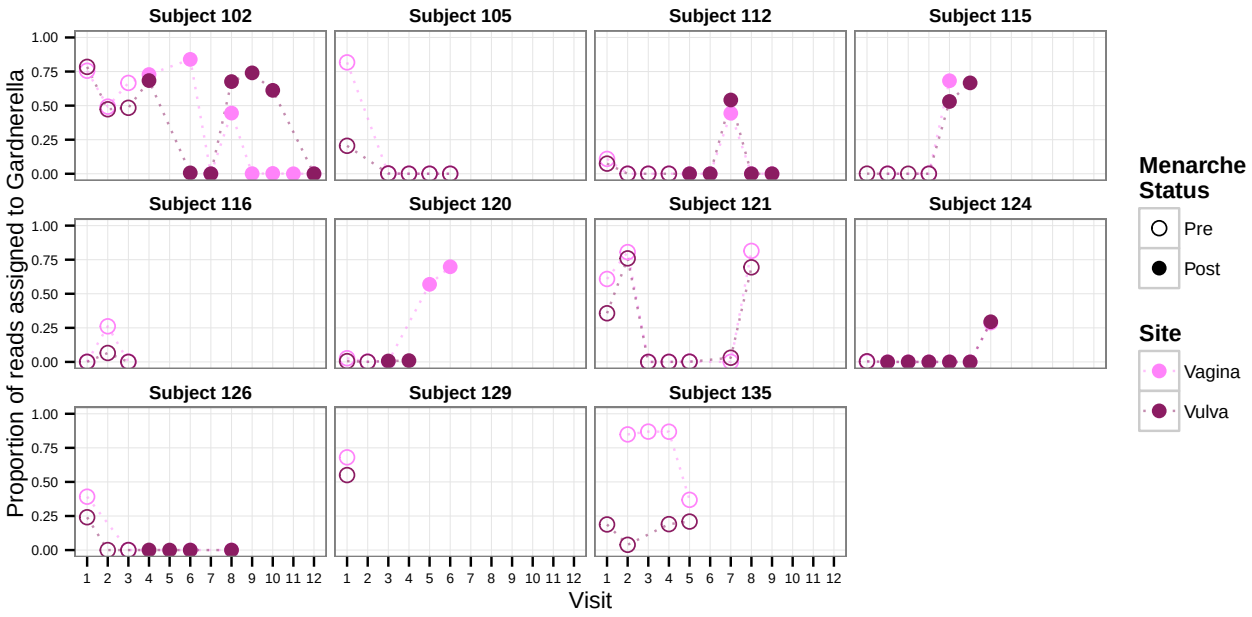
```
## [1] 102 105 112 115 121 124 126 129 135
## 55 Levels: 101 102 103 104 105 106 107 108 109 110 111 112 113 114 ... 235
```

```
meta.girl.gard.10.fullobs <- meta[meta$subject %in% meta.girl.gard.10$subject,]
```

Plot vaginal and vulvar samples of girls with at least 10% *Gardnerella*:

```
gg.girl.gard <- ggplot(meta.girl.gard.10.fullobs,
                       aes(x=visit, y=Gardnerella, group=site, color=site, shape=men.stat))
gg.girl.gard + geom_line(lty=3, alpha=0.5) +
  facet_wrap( ~ gm.pair.fullID, ncol=4) +
  geom_point(size=3) +
  scale_color_manual(values=col.site, name="Site",
                     breaks=c("vag","vul"),
                     labels=c("Vagina", "Vulva")) +
  scale_shape_manual(values=c(16, 1), name="Menarche\nStatus",
                     breaks=c("pre", "post"),
                     labels=c("Pre", "Post")) +
  xlab("Visit") +
  ylab("Proportion of reads assigned to Gardnerella") +
  scale_x_discrete(1:14, name="Visit") +
  ylim(0,1) +
  theme_cust_nominor +
  theme(axis.text=element_text(size=6))
```

```
## geom_path: Each group consist of only one observation. Do you need to adjust the group aesthetic?
```



III. Trends in LAB and vaginal pH with pubertal development

Correlation of Tanner scores

```
## Spearman
cor.sp <- cor(meta[meta$type=="girl" & meta$site=="vag", c("tan.gen.dr","tan.gen.self","tan.br.dr","tan
              use="pairwise.complete.obs", method="spearman")
cor.sp
```

```
##              tan.gen.dr tan.gen.self tan.br.dr tan.br.self
## tan.gen.dr       1.0000       0.6834    0.7273      0.5678
## tan.gen.self     0.6834       1.0000    0.5942      0.7072
## tan.br.dr        0.7273       0.5942    1.0000      0.6336
## tan.br.self      0.5678       0.7072    0.6336      1.0000
```

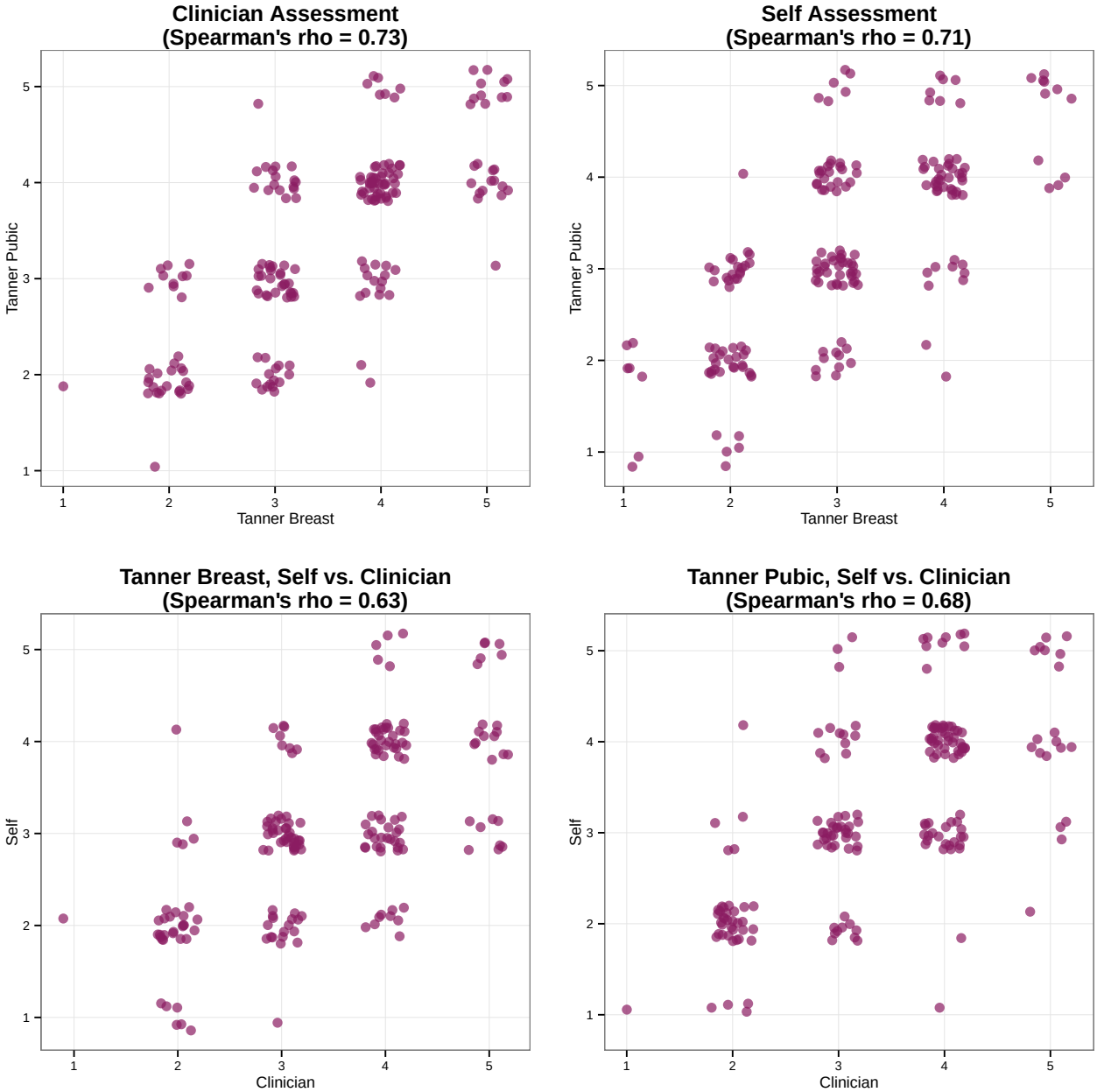
Figure S4. Correlations of Tanner breast and pubic scores according to self and clinician assessment.

Spearman's correlation coefficients were calculated to determine how well breast/pubic scores agreed for both clinician- (upper left) and self-assessment (upper right). Clinician- and self-assessments for Tanner breast (lower left) and Tanner pubic (lower right) scores were also correlated. Correlation coefficients are reported in the header of each plot. Clinician-assessed Tanner breast scores were primarily used for analyses described in the paper.

```
## Tanner pubic vs. Tanner breast (clinician)
gg.tan.dr <- ggplot(meta[meta$type=="girl" & meta$site=="vag",], aes(x=tan.br.dr, y=tan.gen.dr)) +
  geom_jitter(position=position_jitter(width=0.2, height=0.2), color="#8B1C62", size=3, alpha=0.7) +
  ggtitle("Clinician Assessment\n(Spearman's rho = 0.73)") +
  xlab("Tanner Breast") +
  ylab("Tanner Pubic") +
  theme_cust_nominor

## Tanner pubic vs. Tanner breast (self)
gg.tan.self <- ggplot(meta[meta$type=="girl" & meta$site=="vag",], aes(x=tan.br.self, y=tan.gen.self)) +
  geom_jitter(position=position_jitter(width=0.2, height=0.2), color="#8B1C62", size=3, alpha=0.7) +
  ggtitle("Self Assessment\n(Spearman's rho = 0.71)") +
  xlab("Tanner Breast") +
  ylab("Tanner Pubic") +
  theme_cust_nominor

## Tanner breast (self) vs. Tanner breast (dr)
gg.tb.dr.self <- ggplot(meta[meta$type=="girl" & meta$site=="vag",], aes(x=tan.br.dr, y=tan.br.self)) +
  geom_jitter(position=position_jitter(width=0.2, height=0.2), color="#8B1C62", size=3, alpha=0.7) +
  ggtitle("Tanner Breast, Self vs. Clinician\n(Spearman's rho = 0.63)") +
  xlab("Clinician") +
  ylab("Self") +
  theme_cust_nominor

## Tanner pubic (self) vs. Tanner pubic (dr)
gg.tg.dr.self <- ggplot(meta[meta$type=="girl" & meta$site=="vag",], aes(x=tan.gen.dr, y=tan.gen.self)) +
  geom_jitter(position=position_jitter(width=0.2, height=0.2), color="#8B1C62", size=3, alpha=0.7) +
```

```
  ggtitle("Tanner Pubic, Self vs. Clinician\n(Spearman's rho = 0.68)") +
  xlab("Clinician") +
  ylab("Self") +
  theme_cust_nominor

## Uncomment the next line to save as a PDF
# pdf("supplemental/fig-s4-tanner-correlations.pdf", width=10, height=10)
multiplot(gg.tan.dr, gg.tan.self, gg.tb.dr.self, gg.tg.dr.self, layout=matrix(c(1,2,3,4), ncol=2, byrow=
```

```
## Warning: Removed 12 rows containing missing values (geom_point).
## Warning: Removed 9 rows containing missing values (geom_point).
## Warning: Removed 8 rows containing missing values (geom_point).
## Warning: Removed 12 rows containing missing values (geom_point).
```



```
dev.off()
```

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

As the plots above show, the Tanner scores (breast and pubic area, self and clinician) are fairly well correlated. This is not too surprising since the Tanner scores for breast and pubic area were designed to capture simultaneous stages of development. One of the inclusion criteria for participation in the study was Tanner breast of at least stage 2, so it makes intuitive sense to rely the Tanner breast score as a proxy for pubertal development in our study. Additionally, the clinician-assessed Tanner breast score has more complete observations (192 out of 198 vagina samples) than the clinician assessed Tanner pubic score (186/198). Therefore, we'll just use the Tanner breast score for the subsequent analyses.

Get proportions of lactic acid bacteria (LAB)

```
## First we add the proportions of individual Lactobacillus spp., Lactobacillus genus, and lactic acid b
meta$Lactobacillus_crispatus <- t(prop.red["Lactobacillus_crispatus",])
meta$Lactobacillus_gasseri <- t(prop.red["Lactobacillus_gasseri",])
meta$Lactobacillus_iners <- t(prop.red["Lactobacillus_iners",])
meta$Lactobacillus_jensenii <- t(prop.red["Lactobacillus_jensenii",])
meta$Lactobacillus <- colSums(prop.red[grep("Lactobacillus", rownames(prop.red)),])
meta$LAB <- colSums(prop.red[c(6, 28, 43:49, 72:74),]) # 6=Aerococcus, 28=Facklamia, 43:49=Lactobacillu
```

Note: In order to perform linear regression on proportional data (e.g., LAB proportions), the data should be approximately normally distributed. However, our proportions of LAB are heavily skewed with many close to 1 and quite a few close to zero, with far less in the middle. Here we subject the data to a logit transform, log(y/(1-y)), described by Warton and Hui as a preferred method for ecological non-binomial data such as proportions. However, because we have 0's and 1's in our data matrix, these would result in some -Inf and Inf values that cause problems for the linear model. An ad hoc solution to this is to add a nominal value epsilon to the numerator and denominator during the logit transform. Because our data are more skewed toward proportions close to 1 (e.g., check this with hist(meta.vagLAB)), we take as epsilon the difference between 1 and the largest proportion less than 1. This is defined below as eps. For discussion and justification of this approach, see Warton and Hui (2011) The arcsine is asinine: the analysis of proportions in ecology. Ecology 92:3–10.

Logit-transform LAB proportions

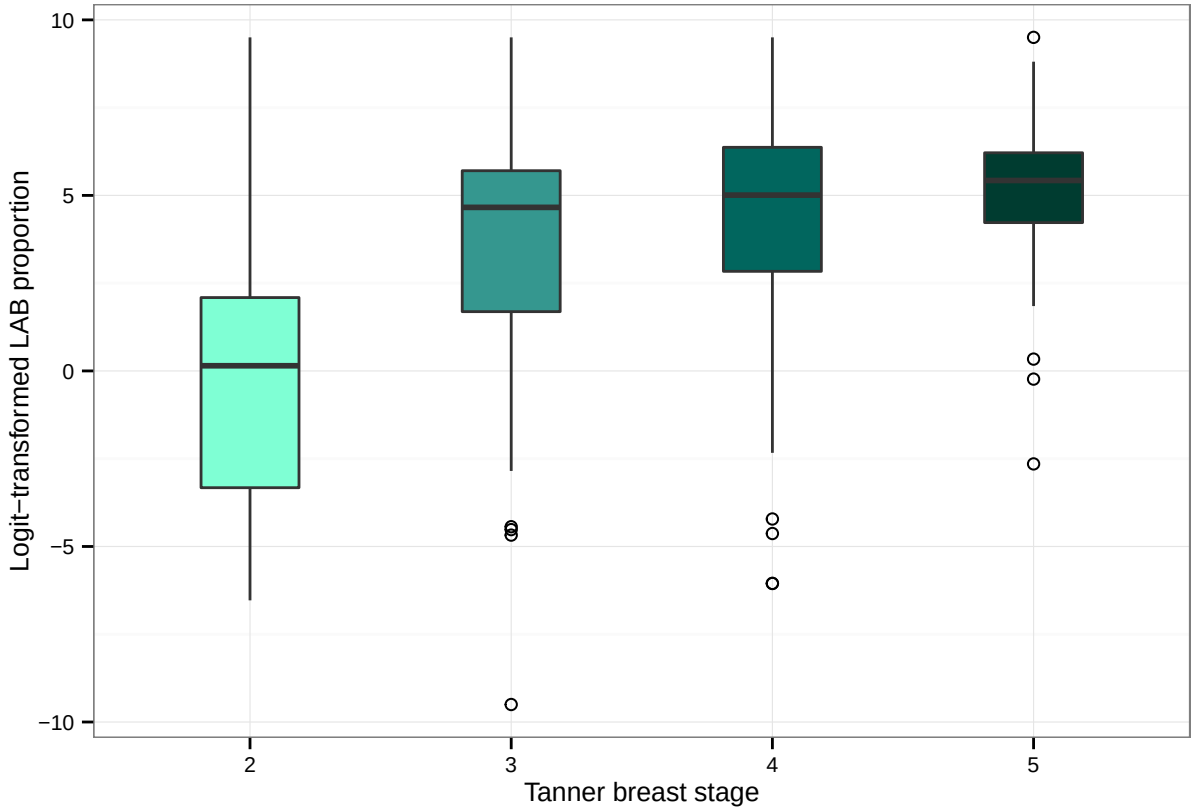
```
## Define modified logit transform so we can specify epsilon adjustment
mod.logit <- function(x, eps) { log((x+eps)/(1-x+eps)) }

## Calculate epsilon:
## Subtract values of LAB from 1
dif.lab <- 1-meta$LAB

## Take the smallest absolute difference greater than zero (~0.000075)
eps.lab <- min(dif.lab[dif.lab > 0])
eps.lab
```

```
## [1] 7.471e-05
```

```
## Add logit-transformed LAB proportions to metadata
meta$LAB.logit <- mod.logit(meta$LAB, eps.lab)
```

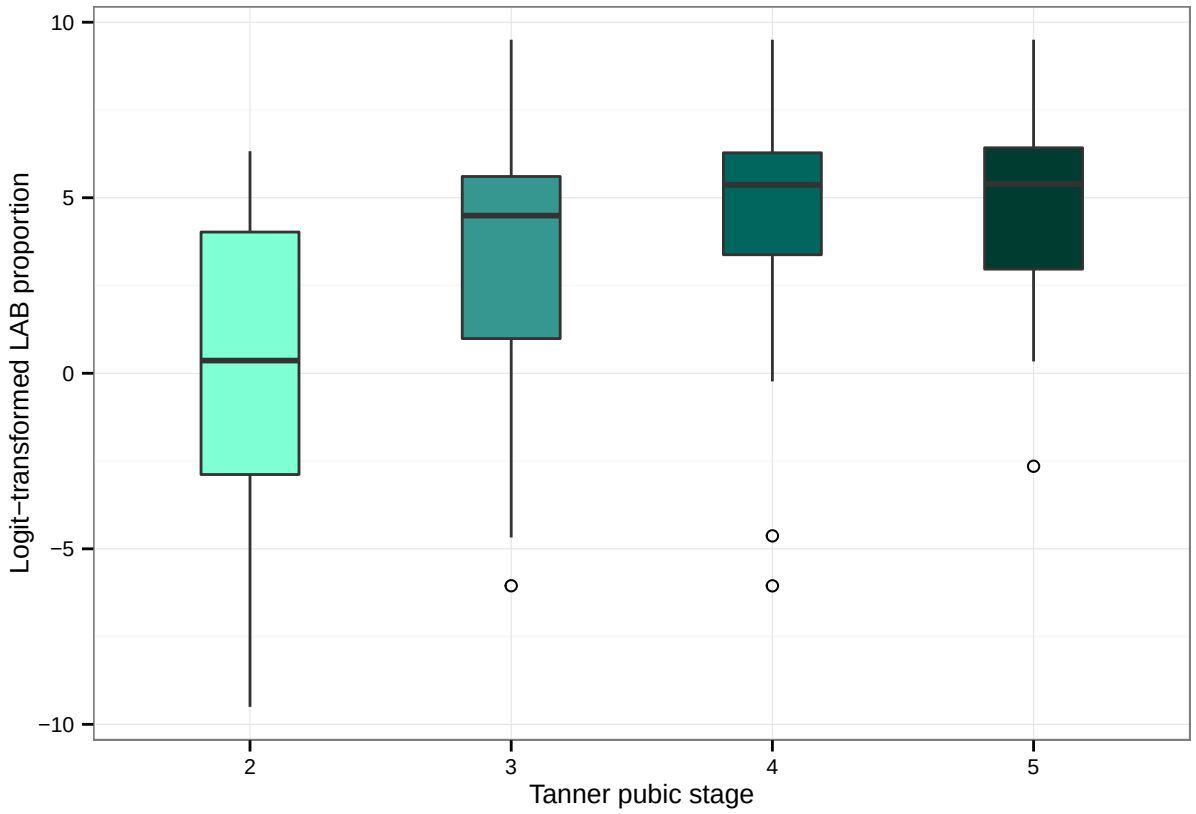
Trends in LAB associated with participant metadata

```
## Boxplot of LAB.logit ~ Tanner breast
gg.lab.logit.tb <- ggplot(subset(meta, type=="girl" & site=="vag" & tan.br.dr %in% c(2,3,4,5)),
                          aes(y=LAB.logit, x=factor(tan.br.dr), fill=factor(tan.br.dr)))
box.lab.logit.tb <- gg.lab.logit.tb +
  geom_boxplot(width=0.5, outlier.shape=1) +
  scale_fill_manual(values=col.tanner[2:5]) +
  guides(fill=FALSE) +
  xlab("Tanner breast stage") +
  ylab("Logit-transformed LAB proportion") +
  theme_cust

print(box.lab.logit.tb)
```

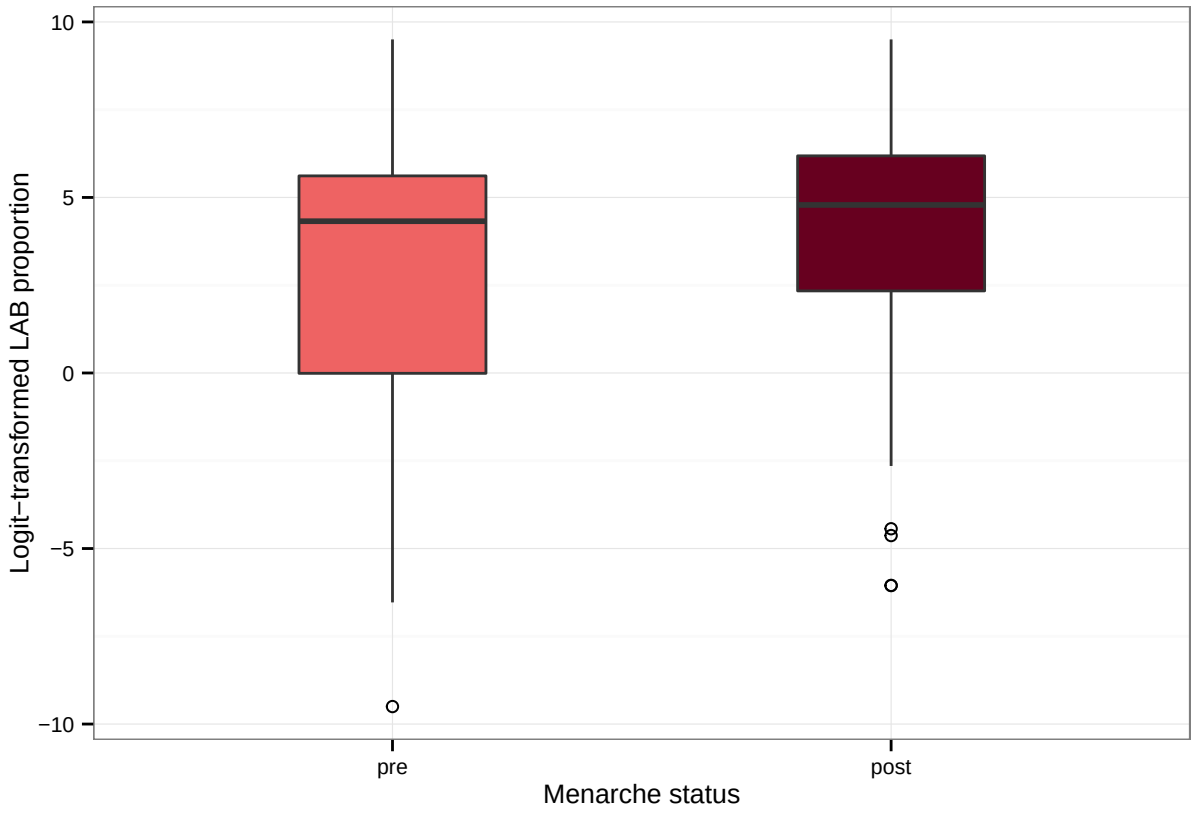



```
## Boxplot of LAB.logit ~ Tanner pubic
gg.lab.logit.tg <- ggplot(subset(meta, type=="girl" & site=="vag" & tan.gen.dr %in% c(2,3,4,5)),
                          aes(y=LAB.logit, x=factor(tan.gen.dr), fill=factor(tan.gen.dr)))
box.lab.logit.tg <- gg.lab.logit.tg +
  geom_boxplot(width=0.5, outlier.shape=1) +
  scale_fill_manual(values=col.tanner[2:5]) +
  guides(fill=FALSE) +
```

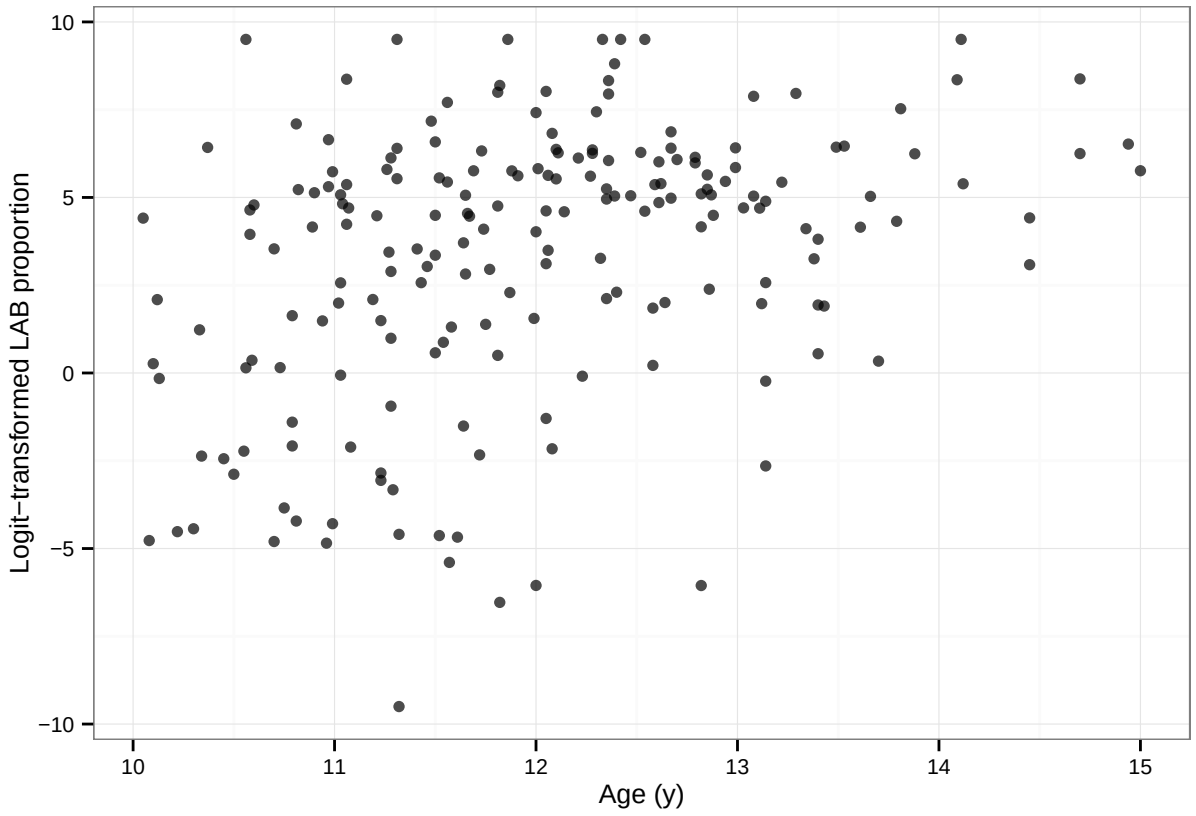
```
  xlab("Tanner pubic stage") +
  ylab("Logit-transformed LAB proportion") +
  theme_cust

print(box.lab.logit.tg)
```



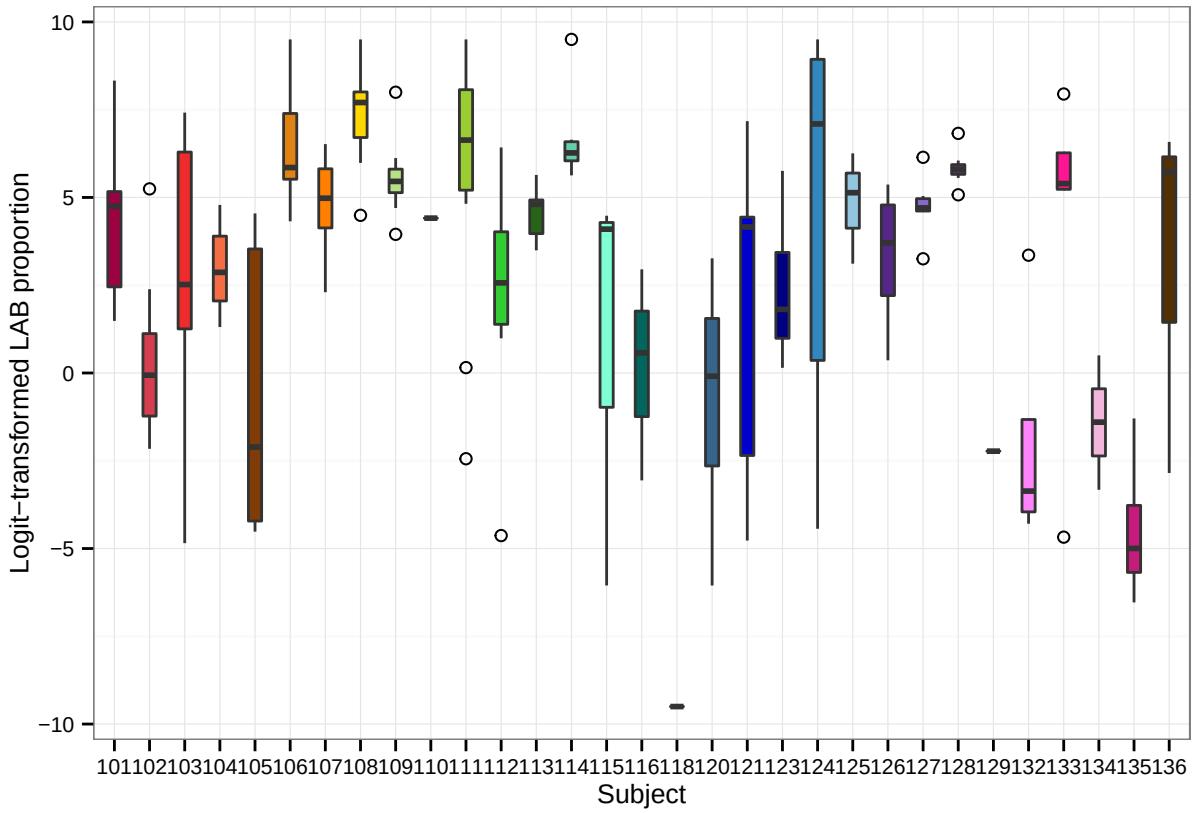

```
## Boxplot of LAB.logit ~ menarche status
gg.lab.logit.ms <- ggplot(subset(meta, type=="girl" & site=="vag" & men.stat %in% c("pre","post")),
                          aes(y=LAB.logit, x=men.stat, fill=men.stat))
box.lab.logit.ms <- gg.lab.logit.ms +
  geom_boxplot(width=0.5, outlier.shape=1) +
  scale_fill_manual(values=col.men.stat[c("pre","post")]) +
  scale_x_discrete(limits=c("pre","post")) +
  guides(fill=FALSE) +
  xlab("Menarche status") +
  ylab("Logit-transformed LAB proportion") +
  theme_cust

print(box.lab.logit.ms)
```




```
## Scatterplot of LAB.logit ~ age (significant in linear model)
gg.lab.logit.age <- ggplot(subset(meta, type=="girl" & site=="vag"),
                           aes(y=LAB.logit, x=age.sampling))
scatter.lab.logit.age <- gg.lab.logit.age +
  geom_point(size=2, alpha=0.7) +
  xlab("Age (y)") +
  ylab("Logit-transformed LAB proportion") +
  theme_cust

print(scatter.lab.logit.age)
```



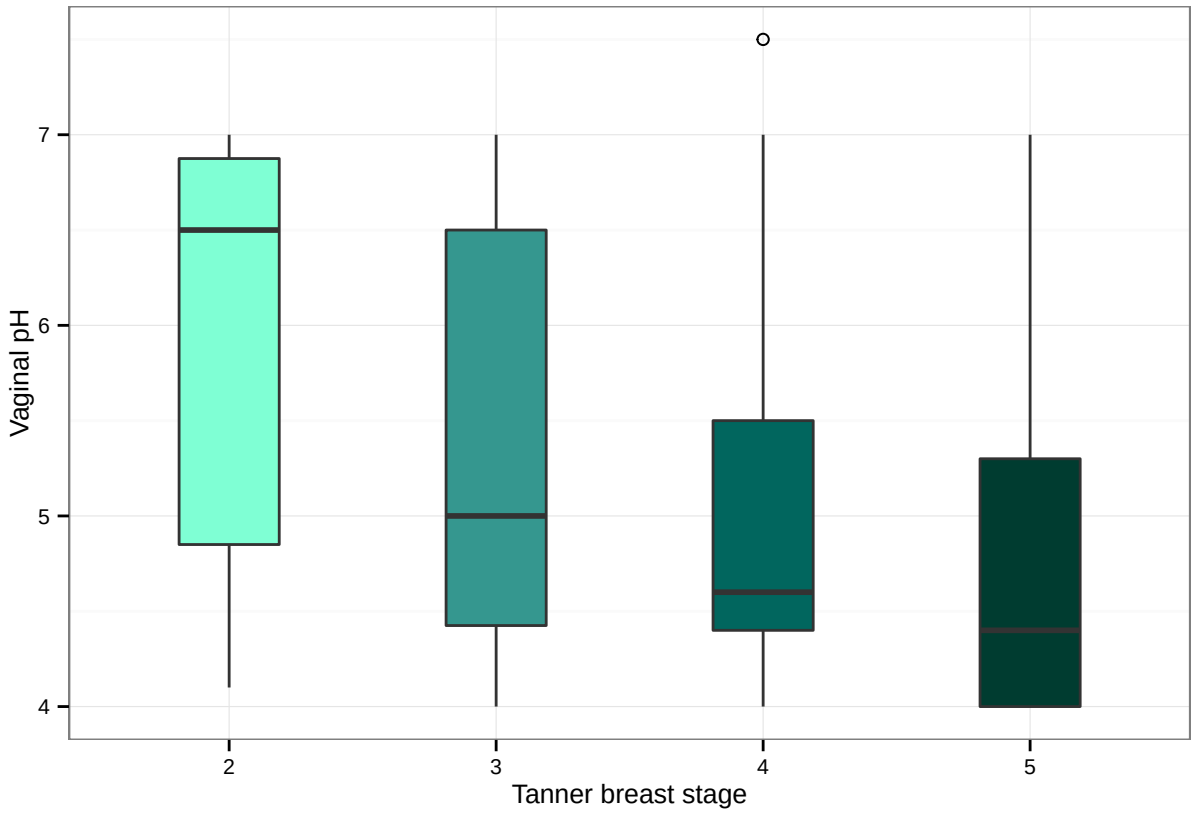

```
## Boxplot of LAB.logit ~ subject (significant in linear model)
gg.lab.logit.sub <- ggplot(subset(meta, type=="girl" & site=="vag"),
                           aes(y=LAB.logit, x=subject, fill=subject))
box.lab.logit.sub <- gg.lab.logit.sub +
  geom_boxplot(width=0.5, outlier.shape=1) +
  scale_fill_manual(values=col.gm.pair[1:31], name="Subject") +
  guides(fill=FALSE) +
  xlab("Subject") +
  ylab("Logit-transformed LAB proportion") +
  theme_cust

print(box.lab.logit.sub)
```



Trends in vaginal pH associated with participant metadata

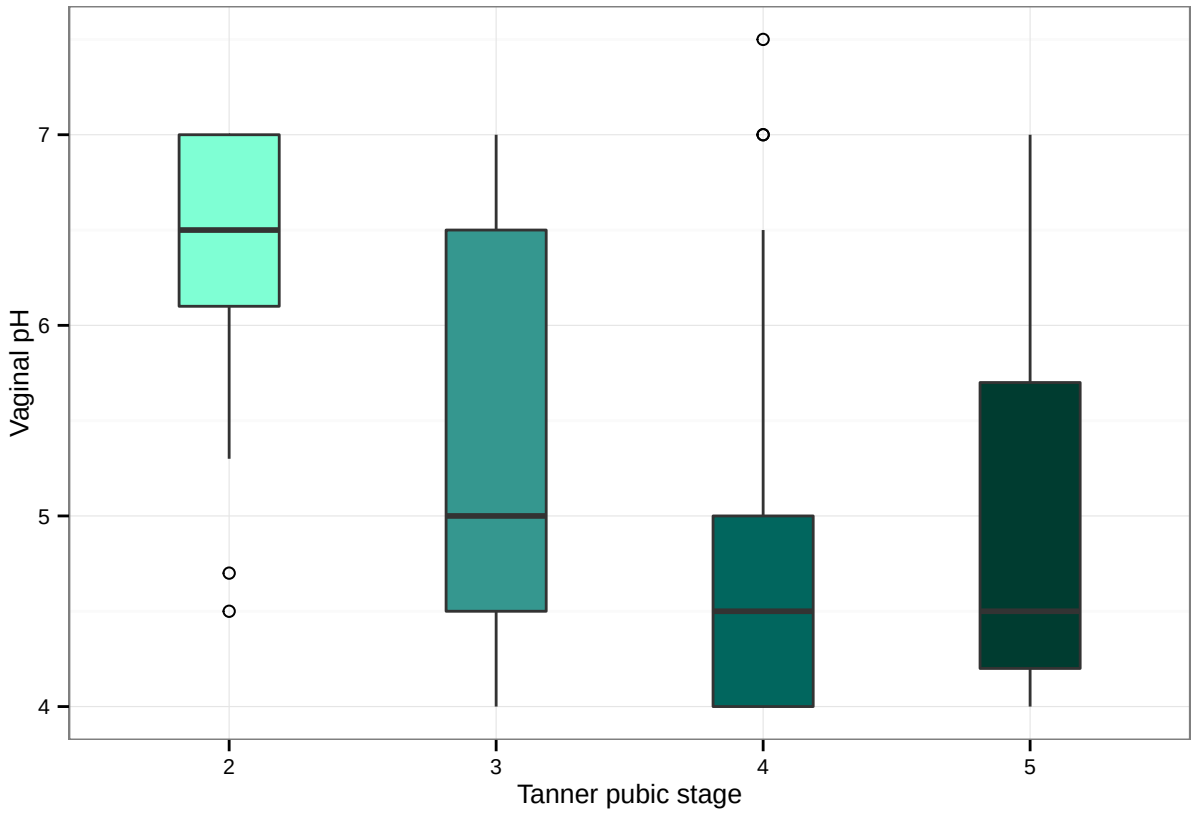
```
## Boxplot of pH ~ Tanner breast
gg.ph.tb <- ggplot(subset(meta, type=="girl" & site=="vag" & tan.br.dr %in% c(2,3,4,5)),
                   aes(y=ph, x=factor(tan.br.dr), fill=factor(tan.br.dr)))
box.ph.tb <- gg.ph.tb +
  geom_boxplot(width=0.5, outlier.shape=1) +
  scale_fill_manual(values=col.tanner[2:5]) +
  guides(fill=FALSE) +
  xlab("Tanner breast stage") +
  ylab("Vaginal pH") +
  theme_cust

print(box.ph.tb)
```

```
## Warning: Removed 62 rows containing non-finite values (stat_boxplot).
```



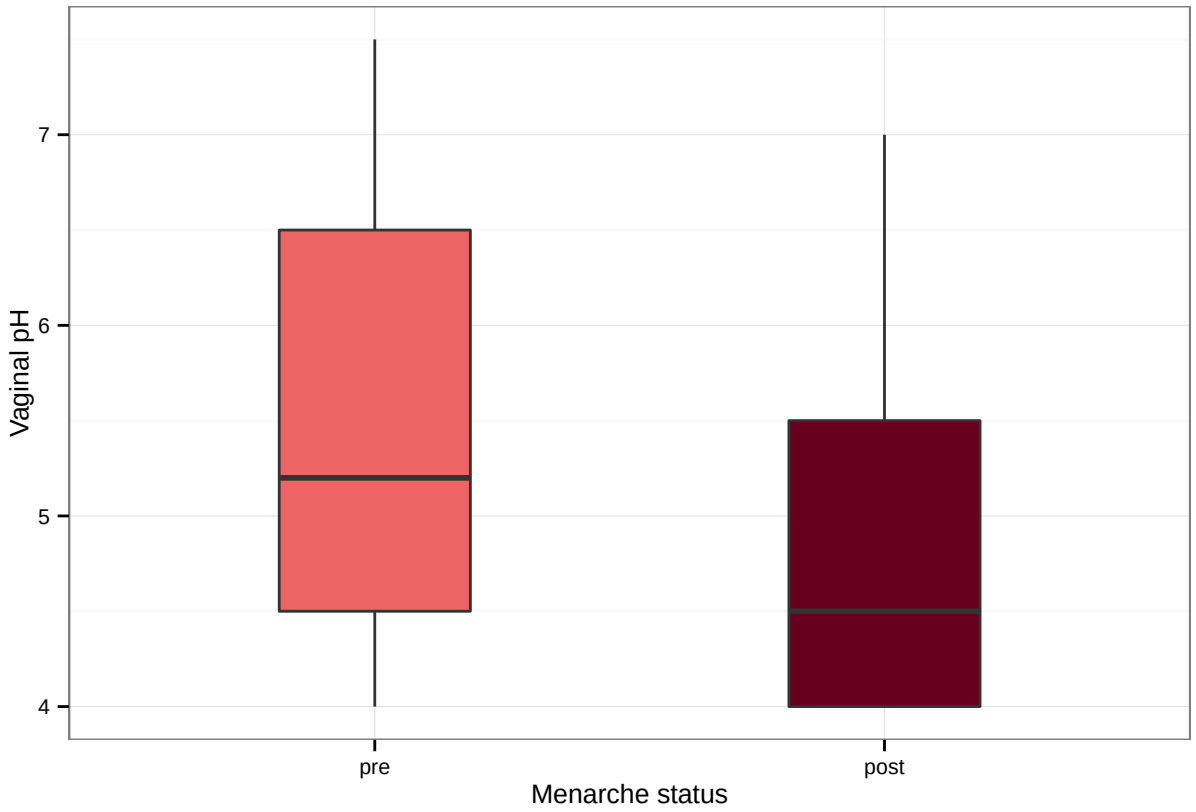

```
## Boxplot of pH ~ Tanner pubic
gg.ph.tg <- ggplot(subset(meta, type=="girl" & site=="vag" & tan.gen.dr %in% c(2,3,4,5)),
                   aes(y=ph, x=factor(tan.gen.dr), fill=factor(tan.gen.dr)))
box.ph.tg <- gg.ph.tg +
  geom_boxplot(width=0.5, outlier.shape=1) +
  scale_fill_manual(values=col.tanner[2:5]) +
  guides(fill=FALSE) +
  xlab("Tanner pubic stage") +
  ylab("Vaginal pH") +
  theme_cust

print(box.ph.tg)
```

```
## Warning: Removed 61 rows containing non-finite values (stat_boxplot).
```



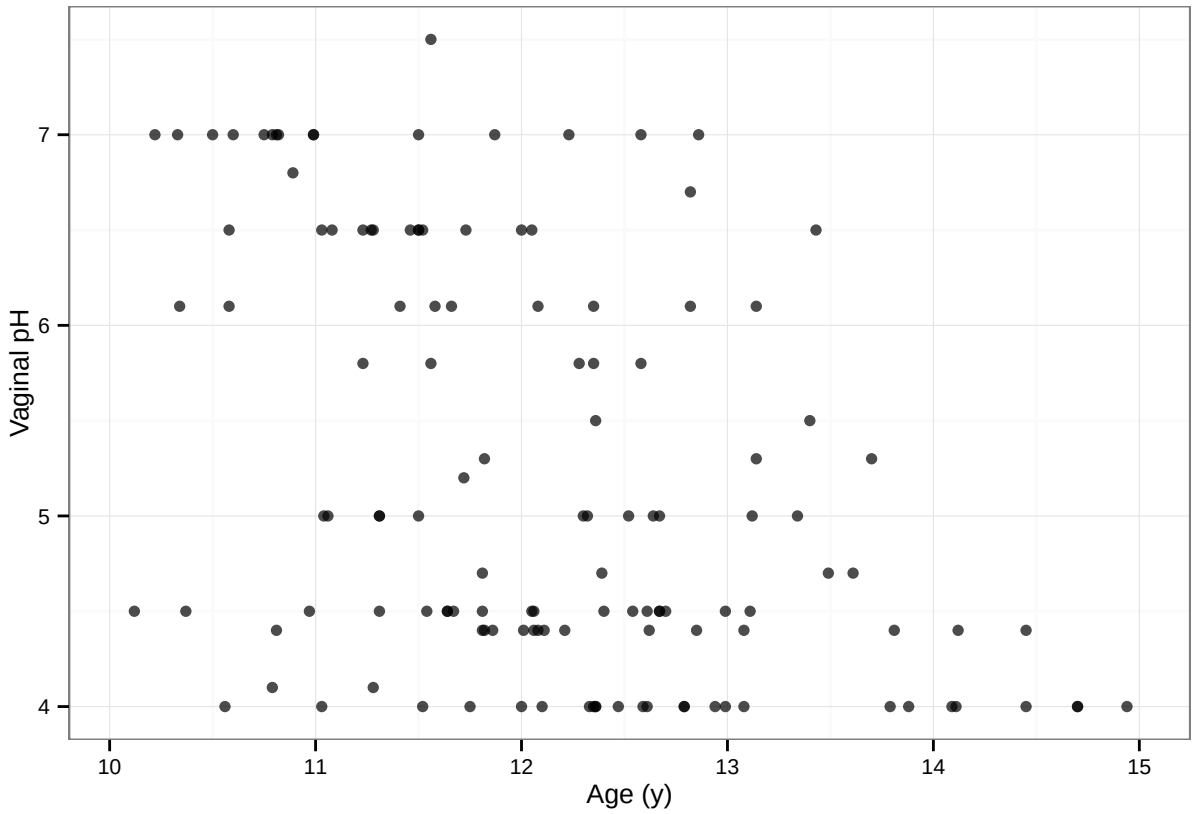

```
## Boxplot of pH ~ menarche status
gg.ph.ms <- ggplot(subset(meta, type=="girl" & site=="vag" & men.stat %in% c("pre","post")),
                   aes(y=ph, x=men.stat, fill=men.stat))
box.ph.ms <- gg.ph.ms +
  geom_boxplot(width=0.5, outlier.shape=1) +
  scale_fill_manual(values=col.men.stat[c(3,2)]) +
  scale_x_discrete(limits=c("pre","post")) +
  guides(fill=FALSE) +
  xlab("Menarche status") +
  ylab("Vaginal pH") +
  theme_cust

print(box.ph.ms)
```

```
## Warning: Removed 67 rows containing non-finite values (stat_boxplot).
```




```
## Scatterplot of pH ~ age (significant in linear model)
gg.ph.age <- ggplot(subset(meta, type=="girl" & site=="vag"),
                    aes(y=ph, x=age.sampling))
scatter.ph.age <- gg.ph.age +
  geom_point(size=2, alpha=0.7) +
  xlab("Age (y)") +
  ylab("Vaginal pH") +
  theme_cust

print(scatter.ph.age)
```

```
## Warning: Removed 68 rows containing missing values (geom_point).
```




```
## Boxplot of ph ~ subject (significant in linear model)
gg.ph.sub <- ggplot(subset(meta, type=="girl" & site=="vag"),
                    aes(y=ph, x=subject, fill=subject))
box.ph.sub <- gg.ph.sub +
  geom_boxplot(width=0.5, outlier.shape=1) +
  scale_fill_manual(values=col.gm.pair[1:31], name="Subject") +
  guides(fill=FALSE) +
  xlab("Subject") +
  ylab("Vaginal pH") +
  theme_cust

print(box.ph.sub)
```

```
## Warning: Removed 68 rows containing non-finite values (stat_boxplot).
```

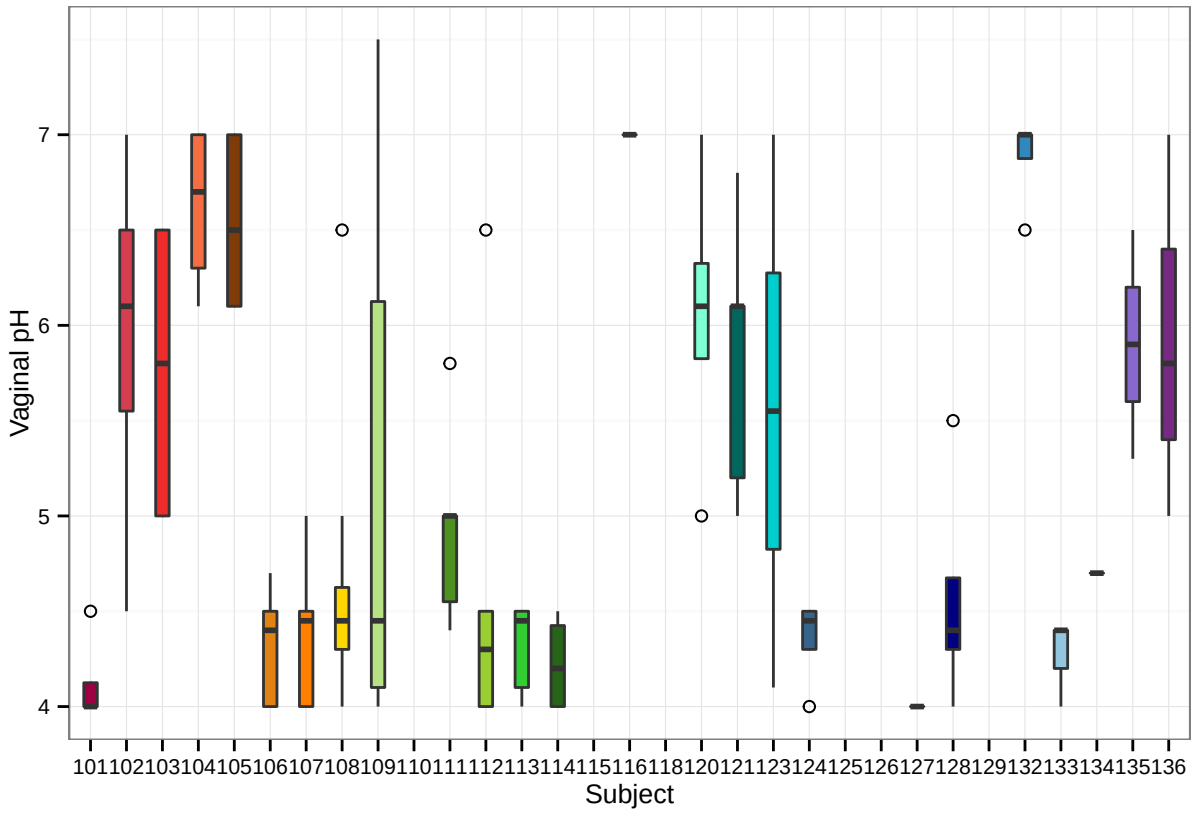


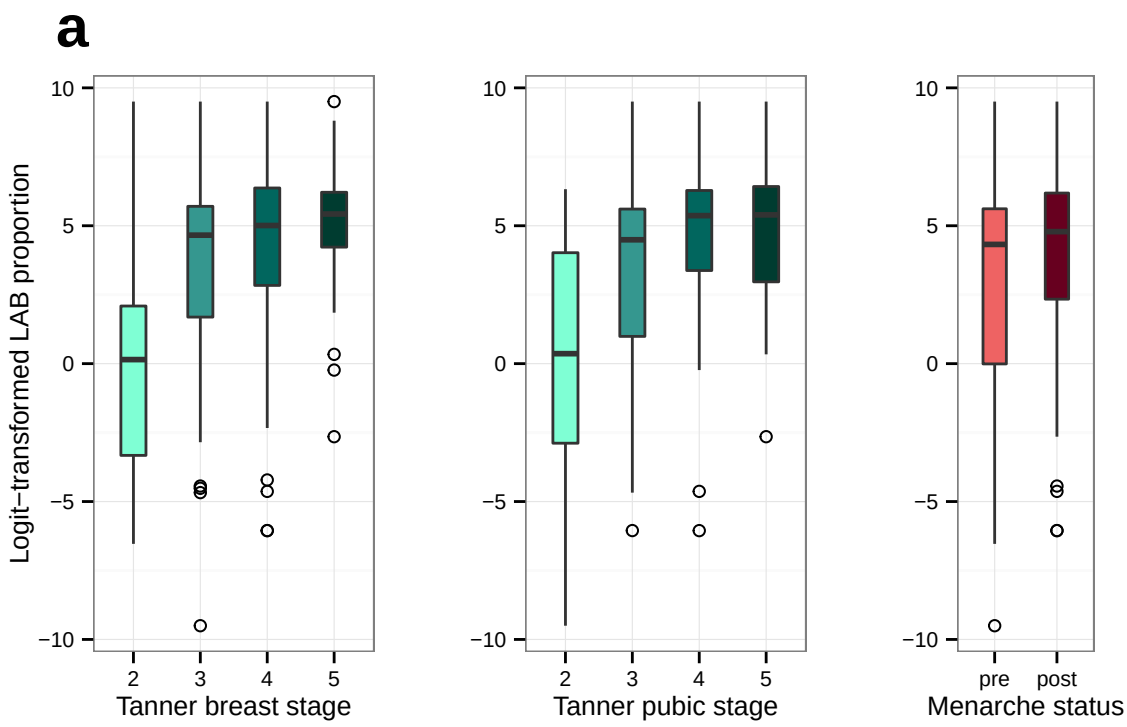
Figure 6. Trends in relative abundance of lactic acid bacteria and vaginal pH in relation to pubertal development and menarche status.

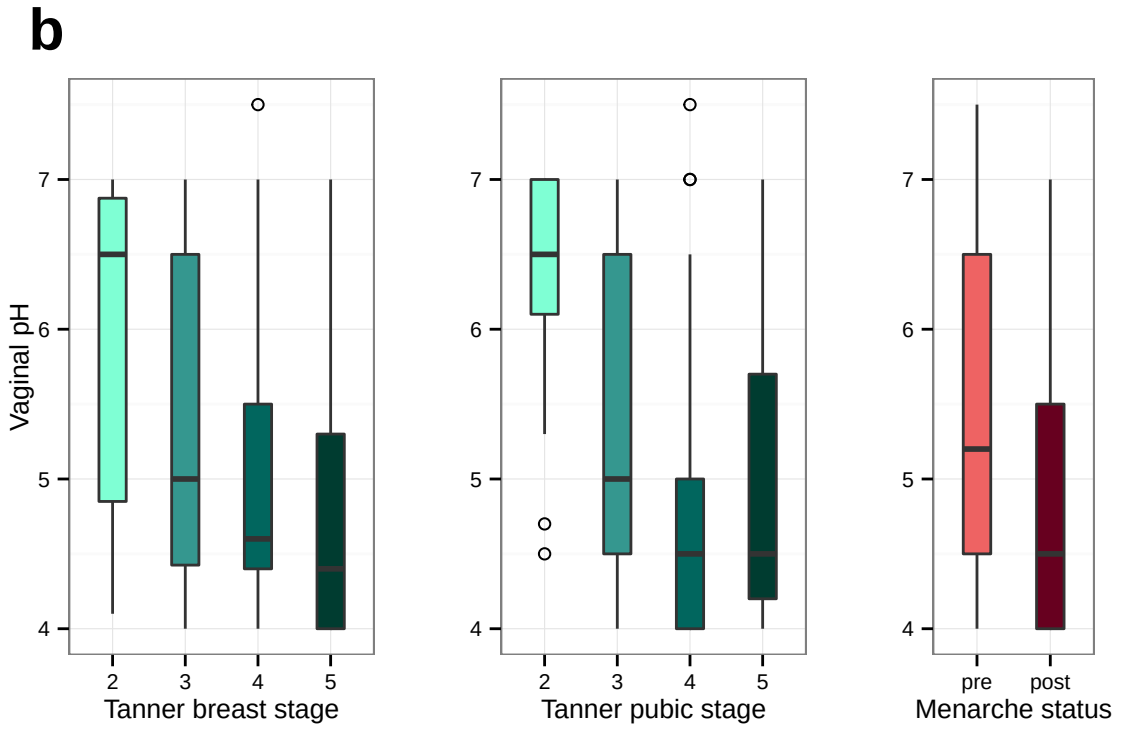
198 vaginal swabs were collected from 31 girls over time. Upper and lower panels show box plots of (a) the logit-transformed proportion of lactic acid bacteria (LAB; includes *Lactobacillus*, *Streptococcus*, *Aerococcus* and *Facklamia*) and (b) vaginal pH. Box plots show the relationship to Tanner breast stage (left column), to Tanner pubic stage (middle) and menarche status (right). In each plot the rectangular box represents the interquartile range, the whiskers represent the upper and lower quartiles, the horizontal line represents the median, and open circles represent outliers.

```
## Make compound figure for manuscript with LAB.logit panels on top, pH on bottom
fig.lab.logit.ph.1 <- box.lab.logit.tb +
  ggtitle("a") +
  theme(plot.title=element_text(size=22, hjust=-0.15, vjust=1.5))
fig.lab.logit.ph.2 <- box.lab.logit.tg +
  ggtitle("") +
  ylab("") +
  theme(plot.title=element_text(size=22, hjust=-0.15, vjust=1.5))
fig.lab.logit.ph.3 <- box.lab.logit.ms +
  ggtitle("") +
  ylab("") +
  theme(plot.title=element_text(size=22, hjust=-0.15, vjust=1.5))
fig.lab.logit.ph.4 <- box.ph.tb +
  ggtitle("b") +
  theme(plot.title=element_text(size=22, hjust=-0.15, vjust=1.5))
fig.lab.logit.ph.5 <- box.ph.tg +
  ggtitle("") +
```

```
  ylab("") +
  theme(plot.title=element_text(size=22, hjust=-0.15, vjust=1.5))
fig.lab.logit.ph.6 <- box.ph.ms +
  ggtitle("") +
  ylab("") +
  theme(plot.title=element_text(size=22, hjust=-0.15, vjust=1.5))

# pdf("figures/fig-6-lab-logit-ph-trends.pdf", width=8, height=6)
multiplot(fig.lab.logit.ph.1, fig.lab.logit.ph.2, fig.lab.logit.ph.3,
          fig.lab.logit.ph.4, fig.lab.logit.ph.5, fig.lab.logit.ph.6,
          layout=matrix(c(1,1,1,2,2,2,3,3,
                          4,4,4,5,5,5,6,6), ncol=8, byrow=TRUE))
```

```
## Warning: Removed 62 rows containing non-finite values (stat_boxplot).
## Warning: Removed 61 rows containing non-finite values (stat_boxplot).
## Warning: Removed 67 rows containing non-finite values (stat_boxplot).
```






```
dev.off()
```

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

Discordance between high LAB and low pH

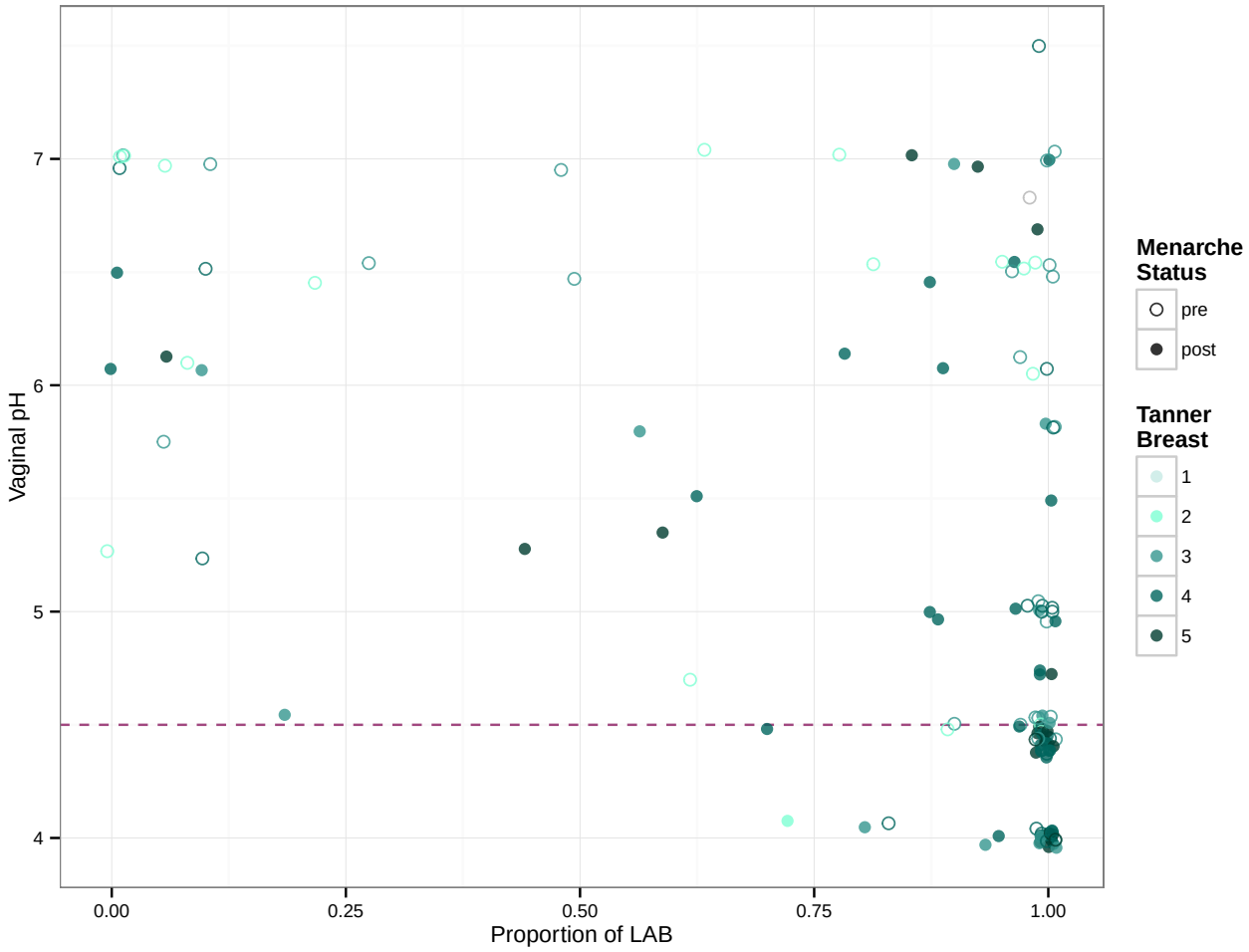
An interesting observation was that not all samples with high proportions of LAB were associated with a low pH. We hypothesized this could be due to lower total bacterial counts in the vaginas of adolescent girls. We performed a coarse test of that hypothesis (see Supplemental File 5, 'adolescent-supp-05.Rmd) but were unable to detect a significant difference in estimated number of 16S rRNA gene copies.

Figure 7. Relationship between proportion of lactic acid bacteria and pH in vaginal samples collected from girls.

197 vaginal microbiota samples from girls are plotted to show the relationship between the proportion of lactic acid bacteria (LAB; includes *Lactobacillus*, *Streptococcus*, *Aerococcus* and *Facklamia*) on the x-axis and vaginal pH on the y-axis. Menarche status is either premenarcheal (open circles) or postmenarcheal (filled circles). Points are color-coded to indicate Tanner breast stage as indicated by the legend at right (NA values are colored gray). Points are slightly jittered to decrease crowding around similar values, but not so much as to distort interpretation of the data. The magenta dashed line at vaginal pH 4.5 represents the upper range of the traditional 'hallmark' healthy vaginal pH of 4.0-4.5.

```
## LAB vs. ph
gg.lab.ph <- ggplot(subset(meta, type=="girl" & site=="vag"),
                    aes(x=LAB, y=ph, col=factor(tan.br.dr), shape=men.stat))

gg.lab.ph + geom_hline(yintercept=4.5, linetype=2, color="#8B1C62", alpha=0.8) +
  geom_jitter(position=position_jitter(width = 0.01, height = 0.05), size=2.5, alpha=0.8) +
  scale_color_manual(values=col.tanner, name="Tanner\nBreast", na.value="gray70") +
  scale_shape_manual(values=c(16,1), name="Menarche\nStatus", breaks=c("pre", "post")) +
  xlab("Proportion of LAB") +
  ylab("Vaginal pH") +
  theme_cust
```

```
## Warning: Removed 68 rows containing missing values (geom_point).
```




```
# Uncomment line below to save as PDF
# ggsave("figures/fig-7-lab-ph-scatterplot.pdf", width=8, height=6, units="in")
```

Cleanup

```
rm(meta.fig5, meta.girl.gard.10, meta.girl.gard.10.fullobs, meta.vag.gard.10,
   meta.vag.gard.10.fullobs, mp.fig5, mp.fig5.lg, mp.s102, mp.s103, mp.s107,
   mp.s109, prop.fig5, prop.fig5.top40, box.lab.logit.ms, box.lab.logit.sub,
   box.lab.logit.tb, box.lab.logit.tg, box.ph.ms, box.ph.sub, box.ph.tb,
   box.ph.tg, dif.lab, eps.lab, fig.lab.logit.ph.1, fig.lab.logit.ph.2,
   fig.lab.logit.ph.3, fig.lab.logit.ph.4, fig.lab.logit.ph.5, fig.lab.logit.ph.6,
   gg.girl.gard, gg.lab.logit.age, gg.lab.logit.ms, gg.lab.logit.sub, gg.lab.logit.tb,
   gg.lab.logit.tg, gg.lab.ph, gg.ph.age, gg.ph.ms, gg.ph.sub, gg.ph.tb, gg.ph.tg,
   gg.tan.dr, gg.tan.self, gg.taxa.bar.102, gg.taxa.bar.103, gg.taxa.bar.107,
   gg.taxa.bar.109, gg.tb.dot.102, gg.tb.dot.103, gg.tb.dot.107, gg.tb.dot.109,
   gg.tb.dr.self, gg.tg.dr.self, scatter.lab.logit.age, scatter.ph.age, taxa.pick40)
```

Save R workspace

This will save the workspace (data) in two separate images: one named with today's date, in case you ever need to restore that version, and another with a non-dated name that can be easily loaded into subsequent analyses.

```
save.image(paste("data-postproc/03-community-dynamics-", Sys.Date(), ".RData", sep=""))
save.image(paste("data-postproc/03-community-dynamics-last-run.RData", sep=""))
```