HW4

I use the paper's data and mutate the data set into one table and then try to reproduce figure 5 in McCallum et al.(2017).

```
library(dplyr)
library(tidyr)
library(directlabels)
library(ggplot2)
```

```
dd1 <- read.csv("POCIS_Raw_McCallum.csv")
dd2 <- read.csv("drugnames.csv")
#head(dd1)
```

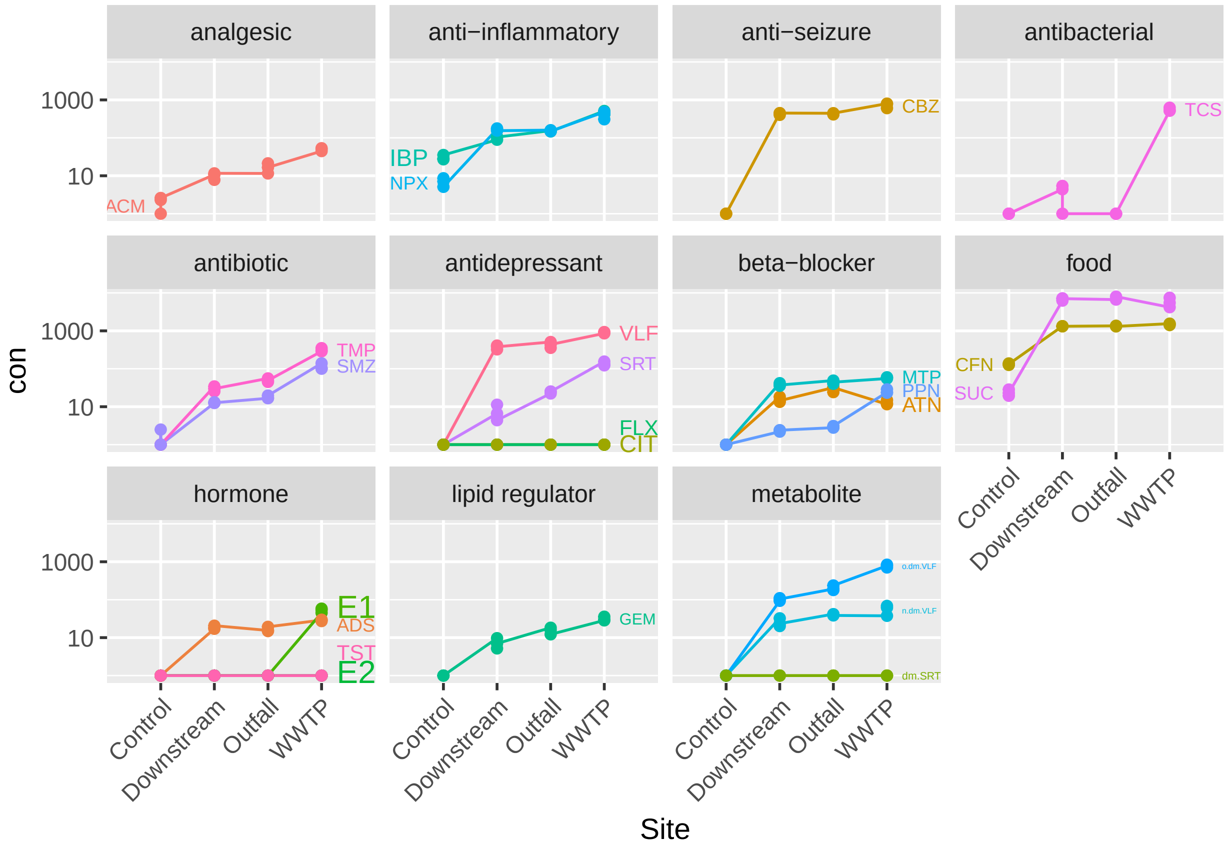
```
dd <- (dd1
       %>% gather(key=abbr,value=con,-c(MetCode,SamplerType,Sample.ID,Site))
       %>% left_join(dd2)
       %>% mutate(con=con+1)
)
```

```
## Joining, by = "abbr"
```

```
## Warning: Column `abbr` joining character vector and factor, coercing into
## character vector
```

```
gg <- (ggplot(data=dd,aes(x=Site,y=con,group=abbr,color=abbr))
    + geom_line()
    + facet_wrap(~drugcat)
    + scale_y_log10()
    + geom_point()
    + theme(axis.text.x = element_text(angle=45,hjust=1))
    + expand_limits(x=c(0,5))
)

nudge_x <- function(x,mid=2,delta=0.2) {
    ifelse(x<mid,x-delta,x+delta)
}
direct.label(gg,method=list(dl.trans(x=nudge_x(x)),"maxvar.qp"))
```



```
## looked at defaultpf.ggplot() to figure out which positioning method
## is being used
```

The graph I made is trying to reproduce the graph in the paper, while after I redo it I found that there are a few differences between the redone work and the original paper.

1. McCallum et al.(2017) combined some of the drug catogories, I am not sure whether there is a biological reason behind, but I think 3x3 graphs indeed looks nicer formatly.

BMB: did you try to do this?

2. The McCallum et al.(2017) made the Sites reversely, for a better intuition of the decreasing pattern for the control group.

BMB: did you try to do this?

3. The transparent geom_point may be better theoretically, but personally I prefer the solid color point, another thing that my graoh annoys me is the directlabel has cover the points sometimes, but I do not know how to separate them a bit. And for the axis.text, I do not know how to move it down a little bit.

BMB: it's perfectly reasonable to have stylistic preferences. For moving axes and labels, see the use of `expand_limits`, `nudge_x` above …

Advantages:

1. Both graphs show the pattern within drug categories and among various types.
2. Both graphs scale the y by log 10 and makes the levels comparable with each other. Friendly with hue and directlabels.

Disadvantages:

It is hard to compare between different categories even though the paper made the comparisons. I consider if we can make them into one row, then we can compare among different catogories easier. However, I am not so sure whether that is the proper thing to do.

BMB:

- it's worth a try, but you'd probably have to lump the categories together more or the panels would be very skinny.
- your lines aren't quite right, they should probably go through the means of each group
- it's probably worth running a spellchecker (RStudio has one built in, I think)
- especially if you're going to do log(1+x), you need a better y-axis label. You could also use `scale_y_continuous(trans="log1p")`
- could you order the groups better? (if it's alphabetical there usually is)
- you should always name your chunks.