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)
```

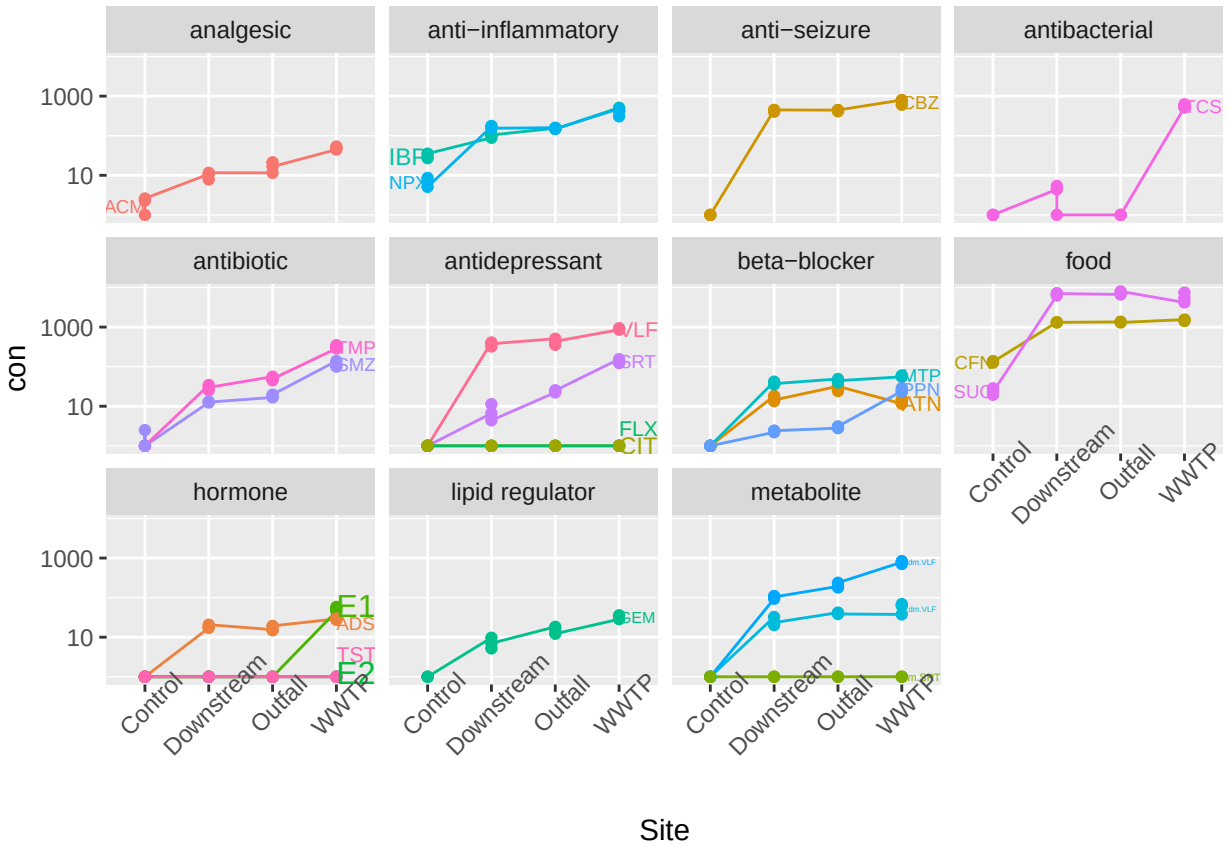
```
dd1 <- read.csv("POCIS_Raw_McCallum.csv")
dd2 <- read.csv("drugnames.csv")
#head(dd1)
```

```
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))
)
direct.label(gg)#+theme(axis.text.x = element_text(angle=45))
```



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 redo work and the original paper.

1. The 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.
2. The McCallum et al.(2017) made the Sites reversely, for a better intuition of the decreasing pattern for the control group.
3. The transparent geom_point may be better theorically, 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 seperate them a bit. And for the axis.text, I do not know how to move it down a little bit.

Advantages:

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

Disadvantages:

It is hard to compare between different catogories 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.

A health science graph reproduced

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) (https://www.sciencedirect.com/science/article/pii/S0166445X16303757?via%3Dihub). Further some details pros and cons are listed.

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

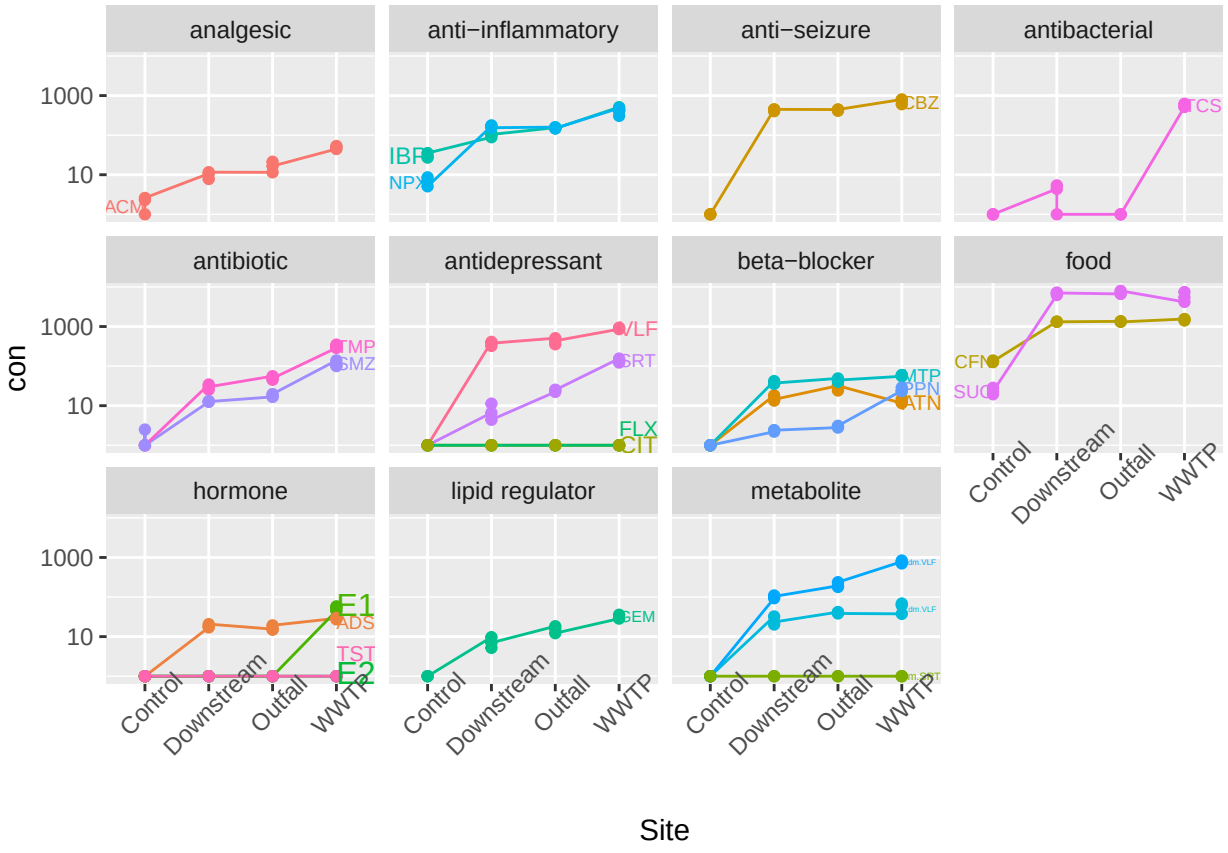
```
dd1 <- read.csv("POCIS_Raw_McCallum.csv")
dd2 <- read.csv("drugnames.csv")
#head(dd1)
```

```
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))
)
direct.label(gg)#+theme(axis.text.x = element_text(angle=45))
```



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 redo work and the original paper.

1. The 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.
2. The McCallum et al.(2017) made the Sites reversely, for a better intuition of the decreasing pattern for the control group.
3. The transparent geom_point may be better theorically, 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 seperate them a bit. And for the axis.text, I do not know how to move it down a little bit.

Advantages:

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

Disadvantages:

It is hard to compare between different catogories 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.