University of Strathclyde Science

BM327 Workshop 2

Identifying UTI Adhesion Factors

Dr Leighton Pritchard and Dr Morgan Feeney

Structure

- Introduction to ggplot2 (R)
- Description of the experiment
- Data analysis (R)

- WebR in your web browser (see MyPlace link)
- https://sipbs-compbiol.github.io/BM327-Workshop-2/

Introduction to ggplot2

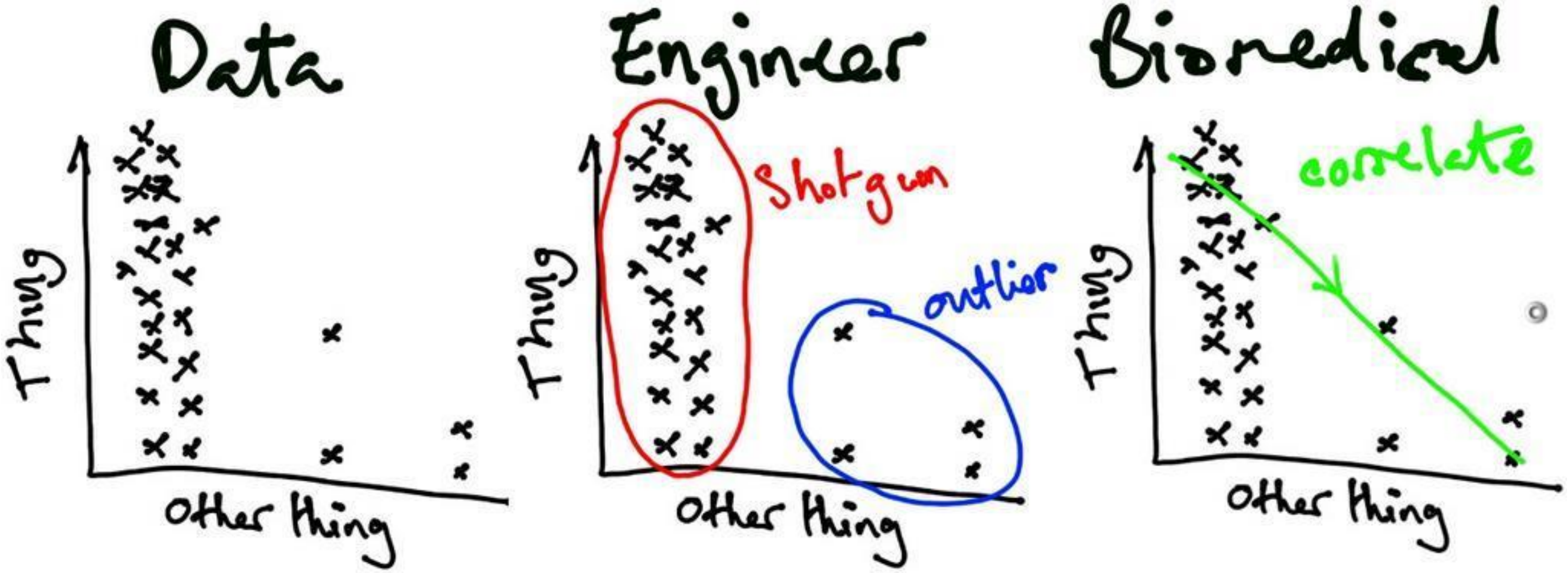
Why ggplot2 and R?

- R is a free (as in beer/chips), widely-used, and robust statistical programming language
- R is excellent for analysis and reproducibility (in science and elsewhere)
  - Separates data from analysis, easy to share/reapply analyses
- R has many useful and advanced statistical tools for experimental/data analysis
- ggplot2 is a powerful, flexible data visualization package in R

WHAT ARE THE ODDS THAT YOU MADE THIS COMPLICATED SPREAD-SHEET WITHOUT ANY CRITICAL ERRORS?
DOES IT MATTER, AS LONG AS IT GIVES ME THE ANSWER I WANT?
IT SHOULD.
BUT ASK YOURSELF IF IT DOES.
Dilbert.com @ScottAdamsSays
© 2016 Scott Adams, Inc. /Dist. by Universal Uclick
1-7-16

Visualisation is critical!

- Data visualisation tells a scientific story
- You need to choose the visualisation that tells the story of the work
  - Being constrained by "available plot types" is limiting
  - ggplot2 allows you to build up the visualisation you need

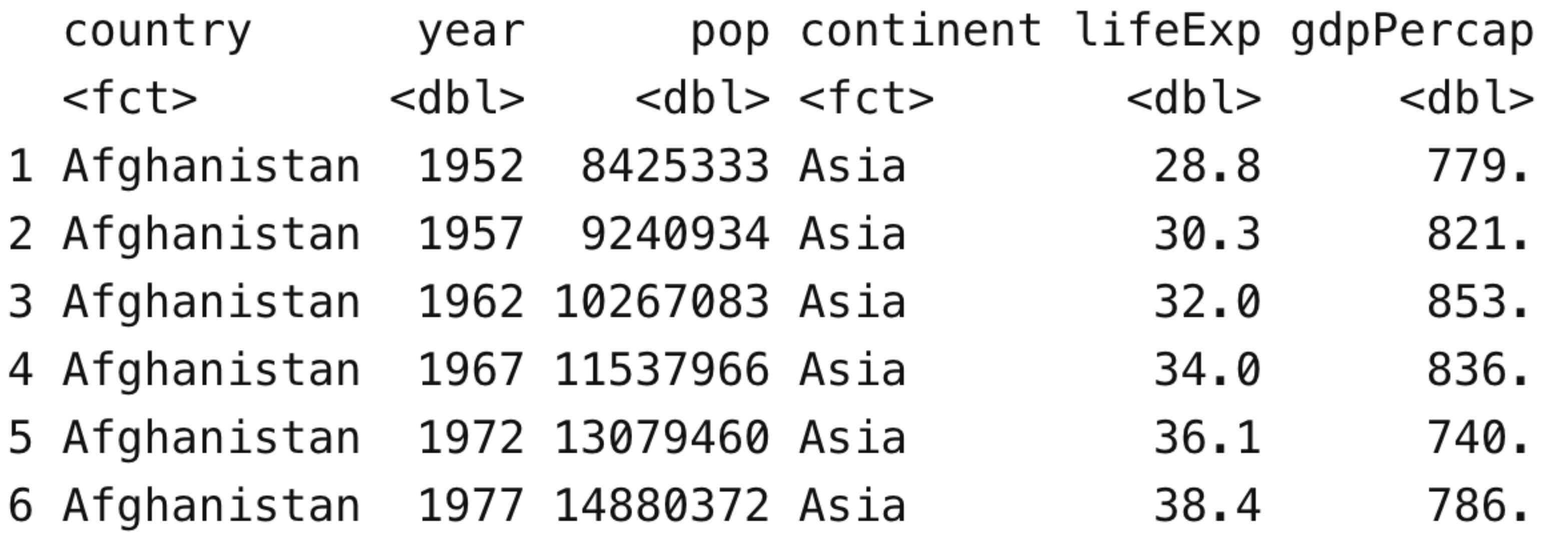


The grammar of graphics

- Separates data from its representation
  - We can make many different possible plots from the same dataset
  - Start by defining the data, and then *layer on* representations of the data
- Build plots from combinations of simple elements
  - Like making a sentence out of adding words together
  - Plots/sentences can be simple or complex, but they should express what you mean
- Data; aesthetics; geoms; layers

What is a plot? (data)

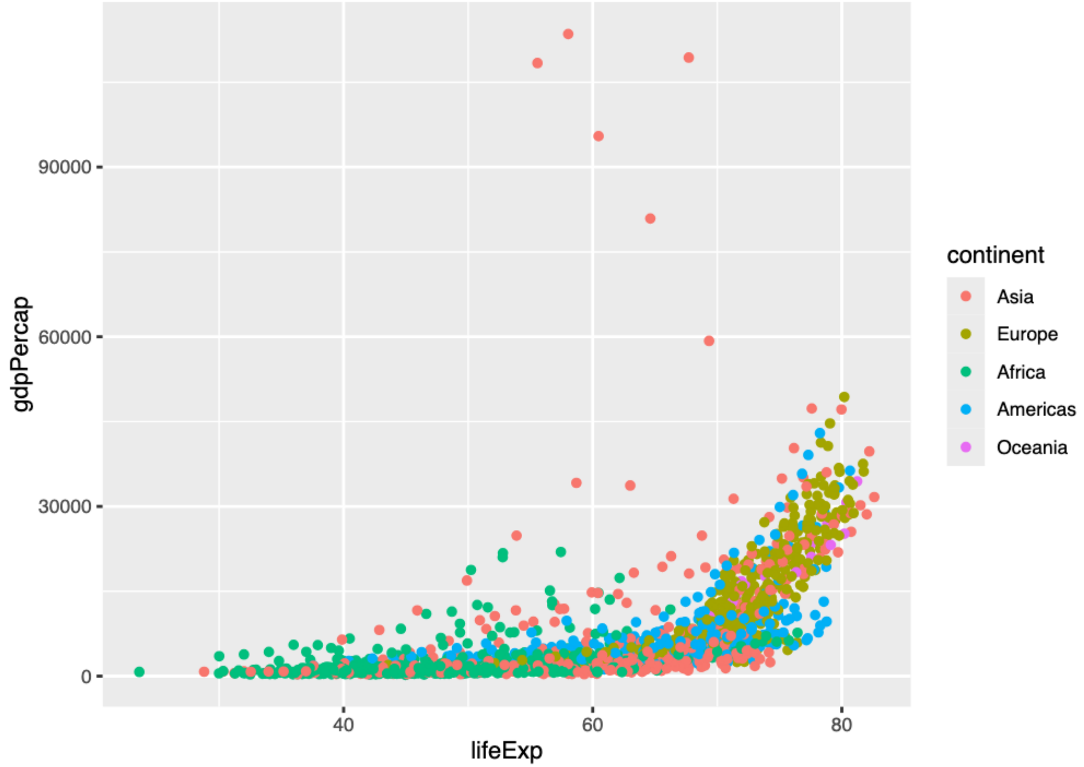
- Your data is usually a table
- One row per observation
- One column per variable
- Each cell is the value of a variable for a particular observation



```
  country      year      pop continent lifeExp gdpPercap
  <fct>       <dbl>    <dbl> <fct>       <dbl>     <dbl>
1 Afghanistan  1952  8425333 Asia         28.8      779.
2 Afghanistan  1957  9240934 Asia         30.3      821.
3 Afghanistan  1962 10267083 Asia         32.0      853.
4 Afghanistan  1967 11537966 Asia         34.0      836.
5 Afghanistan  1972 13079460 Asia         36.1      740.
6 Afghanistan  1977 14880372 Asia         38.4      786.
```

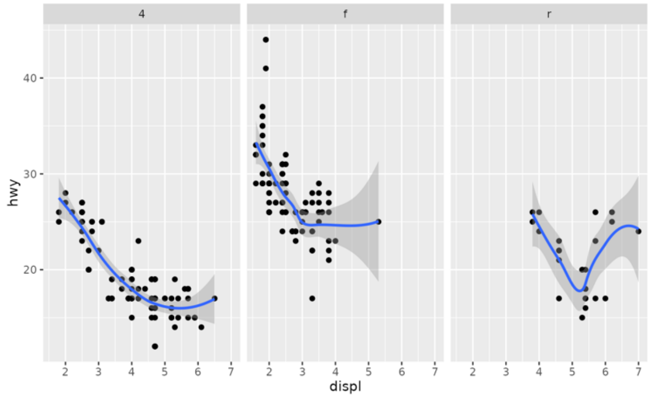
What is a plot (aesthetics)

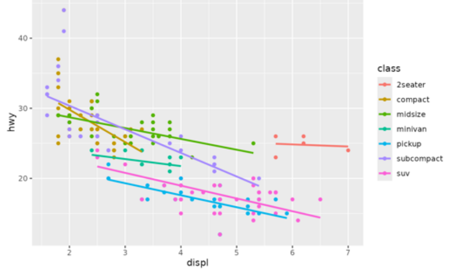
- Each value in the table can potentially be rendered in a plot
- The *aesthetics* of the value determine how it is rendered
  - Shape
  - Size
  - Colour
  - Co-ordinates on the image
- Changing aesthetics changes the plot but not the data
- Many different plots can be made by changing aesthetics alone

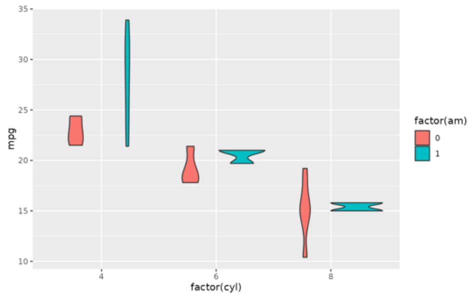


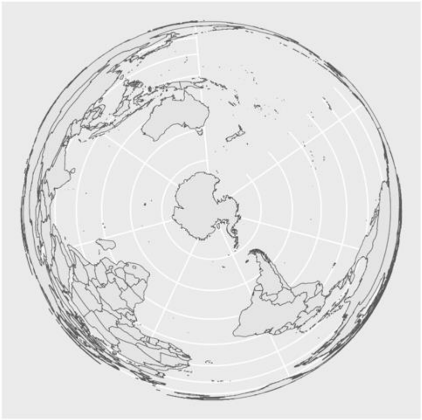
What is a plot? (geoms)

- geoms (“geometries”) are a jargon term in ggplot2
- geoms define the “type” of representation and can be combined
  - Draw as points: scatterplot
  - Draw as lines: line graph
  - Draw as bars: bar chart
  - Draw as box and whisker: boxplot
  - Draw as density plot: KDE/distribution
  - Draw as geographical coordinates: map
  - Draw as vertical density plot: violin plots
  - Draw variability as ribbon: ribbon plots
- The same data/aesthetics can be shown using different geoms



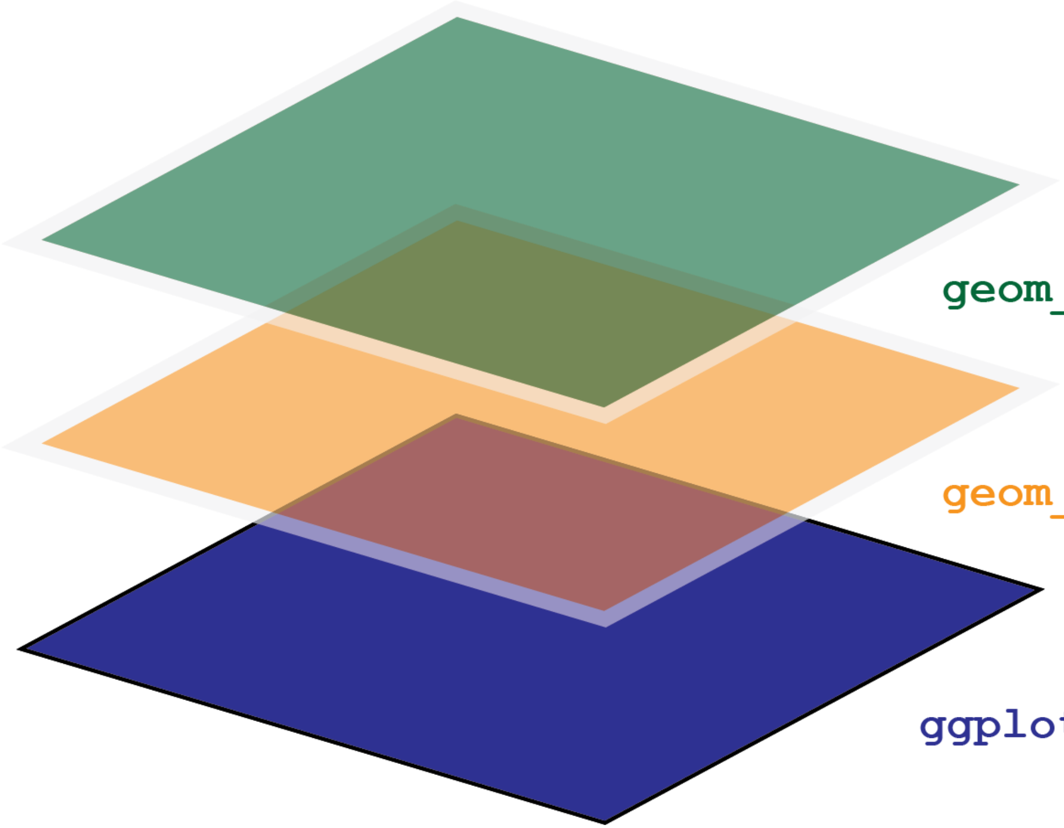






What is a plot? (layers)

- geoms can be combined in layers



```
geom_point(alpha=0.4)

geom_line(aes(group=country))

ggplot(data=gapminder, aes(x=lifeExp, y=gdpPerCapita,
                           colour=continent))
```

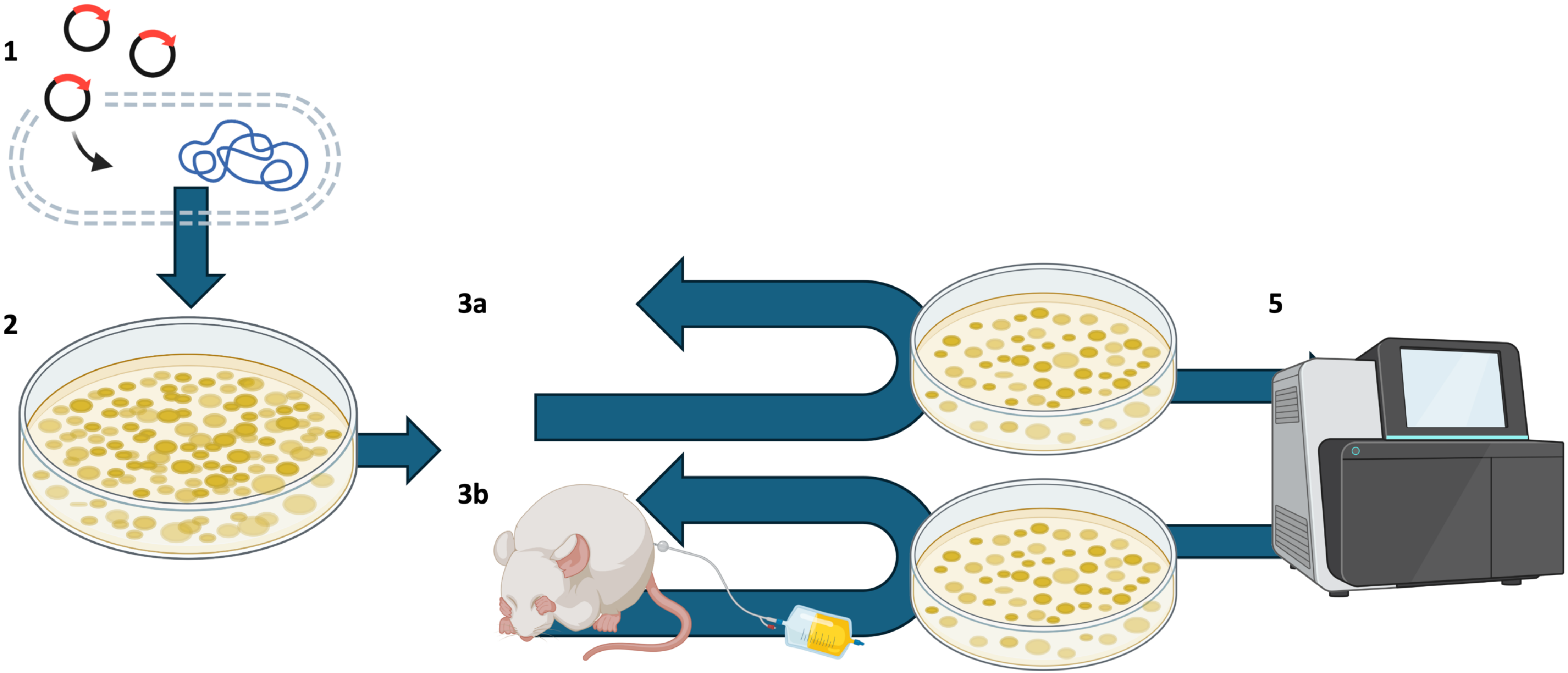
Interactive demo

Let’s work through “The grammar of graphics” on the workshop pages

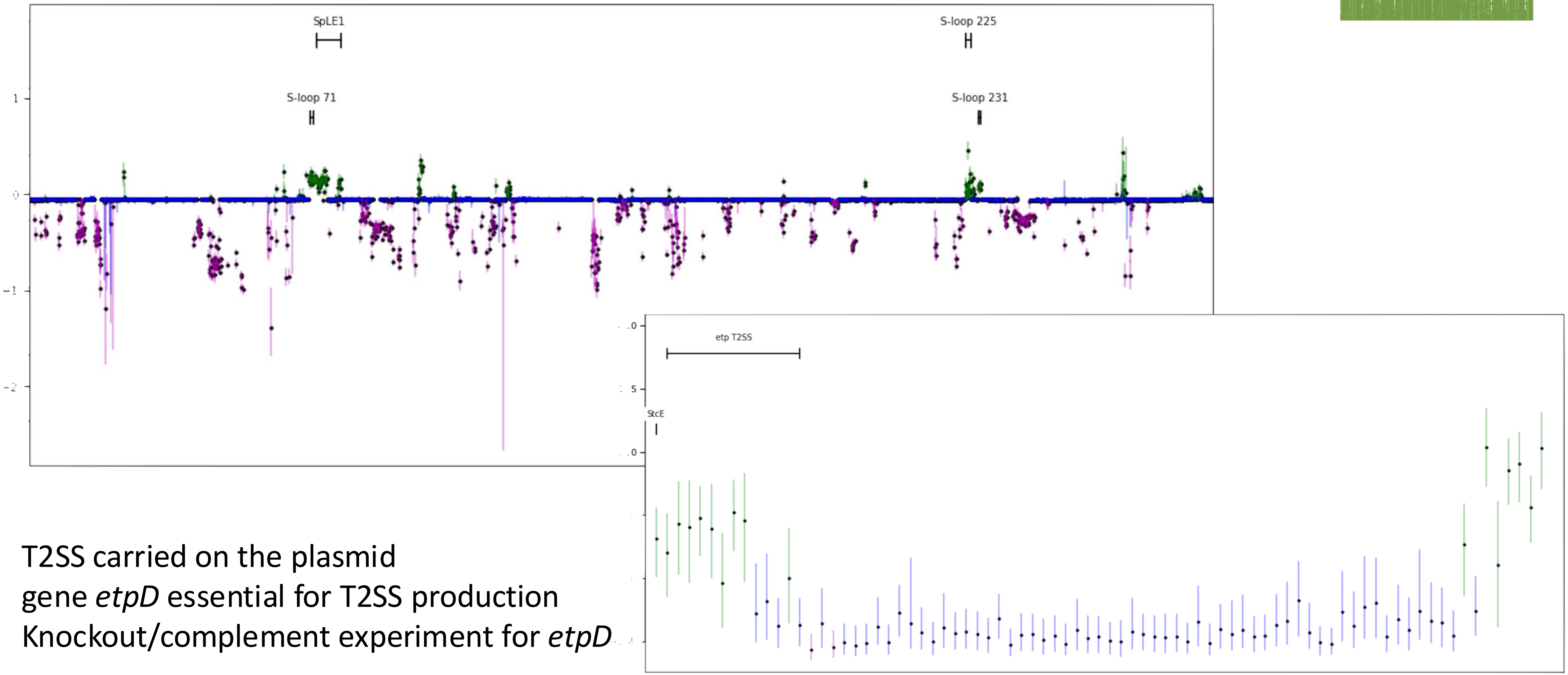
The Experiment

Investigating UTI adhesion

A high-throughput genomic screen



High-throughput results



T2SS carried on the plasmid
gene *etpD* essential for T2SS production
Knockout/complement experiment for *etpD*

Knockout experiment

- (Falkow's) Koch's postulates
  1. The wild-type/control pathogen containing *etpD* must be able to adhere to human tissue/catheter material
  2. The mutant organism lacking only *etpD* must not adhere to human tissue/catheter material
  3. A *complemented* mutant, with *etpD* restored, must be able to adhere to human tissue/catheter material
- We test (catheter material, human tissue sample):
  - Wild-type/control (expected to adhere)
  - etpD knockout (expected not to adhere)
  - etpD knockout with empty plasmid (expected not to adhere)
  - etpD knockout complemented with plasmid carrying etpD (expected to adhere)

Knockout experiment

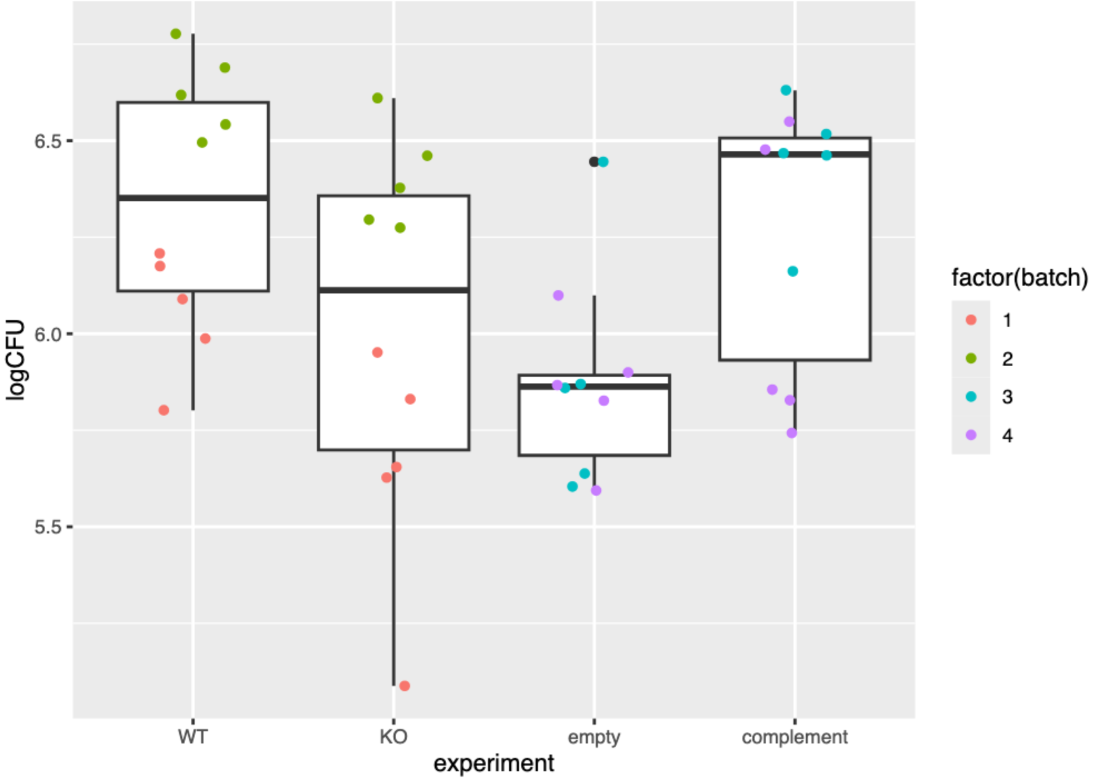
- We introduce to either human tissue or catheter material…
  - Wild-type/control UPEC (expected to adhere)
  - etpD knockout (expected not to adhere)
  - etpD knockout with empty plasmid (expected not to adhere)
  - etpD knockout complemented with plasmid carrying etpD (expected to adhere)
- We thoroughly wash/rinse the material and use serial dilutions to obtain bacterial counts (logCFU)
- High counts imply that bacteria adhered
- Low counts imply that bacteria did not adhere well
- THIS IS AN INDIRECT TEST OF ADHERENCE

The Workshop

What you’ll be doing

Data visualisation

- Use ggplot2 to visualise the experimental results
- Use geom_boxplot() and geom_jitter() geometries
- Colour datapoints by batch
- Obtain plots for catheter and human tissue
- What do you notice?



Statistical modelling

- This may well be new to you
- A different philosophy to null hypothesis significance testing (NHST)
  - (things like t-tests, ANOVA, etc.)
- We’ll use **linear modelling** (simple to do in R)
- We explicitly, simultaneously, and quantitatively estimate the effect of each intervention, relative to the wild-type/control:
  - etpD knockout
  - addition of empty plasmid
  - complementation
  - any interference effects (e.g. batches of experiments run at different times/with different media/by different people)

A high-level view of linear modelling

- We are measuring some kind of outcome
  - Here, we measure logCFU bacterial recovery
- We assume that the measured value depends (“~”) on some influence
- The measured logCFU for the wild-type UPEC depends on us using the wild type

logCFU ~ wildtype

A high-level view of linear modelling

- In reality, there is some variation in measurement
  - e.g. wind on the balance, slight differences in growth time
- We assume these variations are random, and represent them as ε
- Linear modelling lets us “subtract” these random effects and estimate the actual influence of “wildtype”

$$\text{logCFU} \sim \text{wildtype} + \varepsilon$$

A high-level view of linear modelling

- We can account for multiple influences by adding further terms into the equation
- When considering the knockout strain for example, there are two influences
  - Recovery appropriate for the wildtype
  - A change in recovery due to the ΔetpD knockout (expected to be negative)
- We assume that we can add these
- Linear modelling estimates the influences of both wildtype and ΔetpD simultaneously

$$\text{logCFU} \sim \text{wildtype} + \Delta\text{etpD} + \varepsilon$$

A high-level view of linear modelling

- We can extend this to all of the experimental factors.
- There are two influences
  - Recovery appropriate for the wildtype
  - A change in recovery due to the ΔetpD knockout (expected to be negative)
  - A change in recovery due to presence of the plasmid vector (expected to be 0)
  - A change in recovery due to presence of the complement (expected to be positive)
- Linear modelling estimates the influences of all of these simultaneously

logCFU ~ wildtype + ΔetpD + vector + complement + ε

Accounting for batch effects

- This experiment is subject to batch effects
- These become an extra term in the equation for linear modelling
- If there are four batches we write this as batch_i to mean the appropriate one of $\{\text{batch}_1, \text{batch}_2, \text{batch}_3, \text{batch}_4\}$ and add it to the equation
- This is a “linear mixed effects model”, and subtracts out the influence due to each individual batch to give better estimates of experimental factors

$$\text{logCFU} \sim \text{wildtype} + \Delta\text{etpD} + \text{vector} + \text{complement} + \text{batch}_i + \varepsilon$$

Linear modelling in R

- Linear model

    -
      ```
      tissue_model <- lm(logCFU ~ KO + empty + complement,
                         data=tissue)
      ```
- Mixed effects model

    -
      ```
      tissue_mixed_model <- lmer(logCFU ~ KO + empty + complement +
                                 (1 | batch), data=tissue)
      ```

Interactive Demo

Let’s work through the workshop pages