The Seven Principles of Big Data Biology

Big Data Biology (BIO00047I) | January 2019
Workshop material:
http://www-users.york.ac.uk/~dj757/BIO00047I/BIO00047I_index.html

1. **What is the biological question?**
   This is the most important principle by far! Keep in mind what you want to find out. This will affect how the data is collected. It also affects how you analyse the data, how you show it with plots, and what statistical tests you do.
2. **Know your data.**
   Get to know your data like a friend, by asking it questions. Where did it come from? How far can you trust it? How many data points are there? Is the data normally distributed (like a bell curve), or are there many small values and only a few very large ones. Or, is the data categories ('categorical data').
   Hint: plot first, then do stats.
3. **Filter out the bad stuff.**
   Don't let bad data lead you to bad conclusions. Garbage in means garbage out. We'll teach you what the bad stuff is, don't worry. **Hint:** plot again after filtering.
4. **Statistical tests are powerful.**
   Statistical tests are wonderful, powerful ways of describing and understanding data. But we need to choose the test carefully to suit the data, and to answer a specific question about the data.
5. **Be careful with statistics.**
   Statistics are powerful, but they can mislead us. For example, statistics alone would suggest that eating chocolate might get you a Nobel Prize, *see here*.
   The solution is: *Know your data* (see above).
6. **If you conduct multiple tests, you need multiple test corrections.**
   For example, if you had a 100-sided die, what's your chances of rolling a 3? What about if you had 100,000 monkeys* and each of them rolled a 100-sided die? What are your chances of getting at least one 3 then?
   **We don't have ethical approval for this experiment.*
7. **Present data honestly and clearly.**
   A picture is worth a thousand words*. Make plots clear and simple. Use colour. Label axes. Please don't make 3D plots. Big data *needs* plots. We love them. OK, that's enough about plots.
   For the purposes of assessment, a picture is **not worth a thousand words.*