Summary: Introduction to Statistics and Experimental Design

Reuben Thomas
bioinformatics@gladstone.ucsf.edu

February 17, 2021

1. *Ideal when presenting evidence for a new scientific claim:* We want to be able to convince the most rational but sceptical person or the one for whom it will take a lot to convince of anything new.

2. Data from biological experiments are inherently noisy. However, it is reasonable to model these noisy data as coming from a given (unknown) probability distribution or what is known as the **data generating distribution**

3. The generality of the claim we would like to (or can) make determines (or is determined) by the **target population** (e.g., all mice of a given strain, all children below 5 years) we have access to or sample from.

4. Statistical theory through the use of **p-values** and **confidence intervals** allows us to make claims about the entire **target population** despite only sampling a typically small subset of.

5. Implicit in all **p-values** and **confidence intervals** are **parameters of interest** (e.g., difference in mean levels of a given marker, fold-change, odds ratio, hazard ratio)

6. The generality of the claims to the entire **target population** using **p-values** and **confidence intervals** depends on whether the computation of these took into account the sample-to-sample variability of the responses of interest. This means that you need to ensure that you have a **"reasonable" number of replicates** from the target population.

7. Very, very important to make sure that your **parameter of interest** is not **confounded** by other stuff - batch of processing, gender etc.

8. If the claim you want to make is across multiple genes, markers then make sure you correct for **multiple testing**.

9. Before starting with any wet-lab work, have a hypothesis and an experimental plan that ensures one can reasonably rule out **chance**, **confounding** and **bias** for whatever claim you make afterwards. *If you are not sure then talk to us in the Bioinformatics Core*

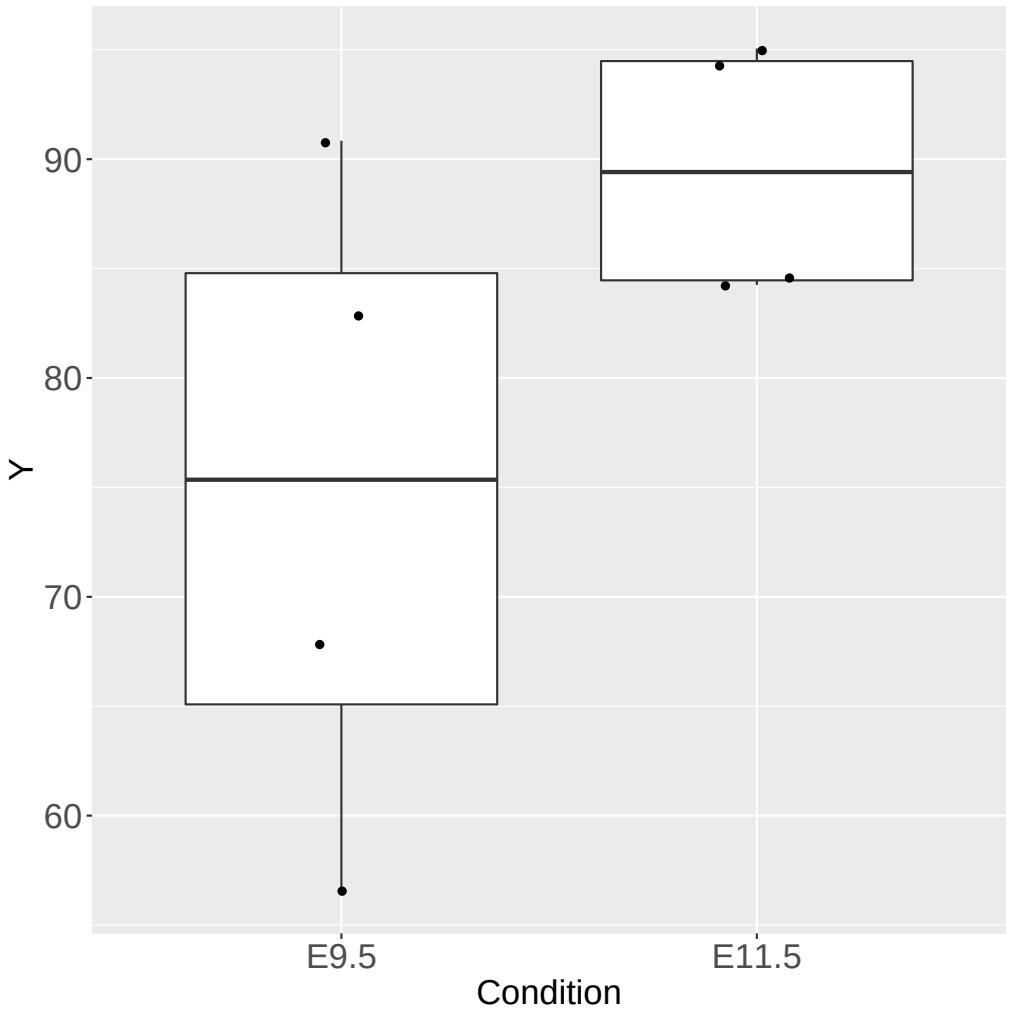

Figure 1:Distribution of expression of gene X across 4 replicates over two developmental time-points

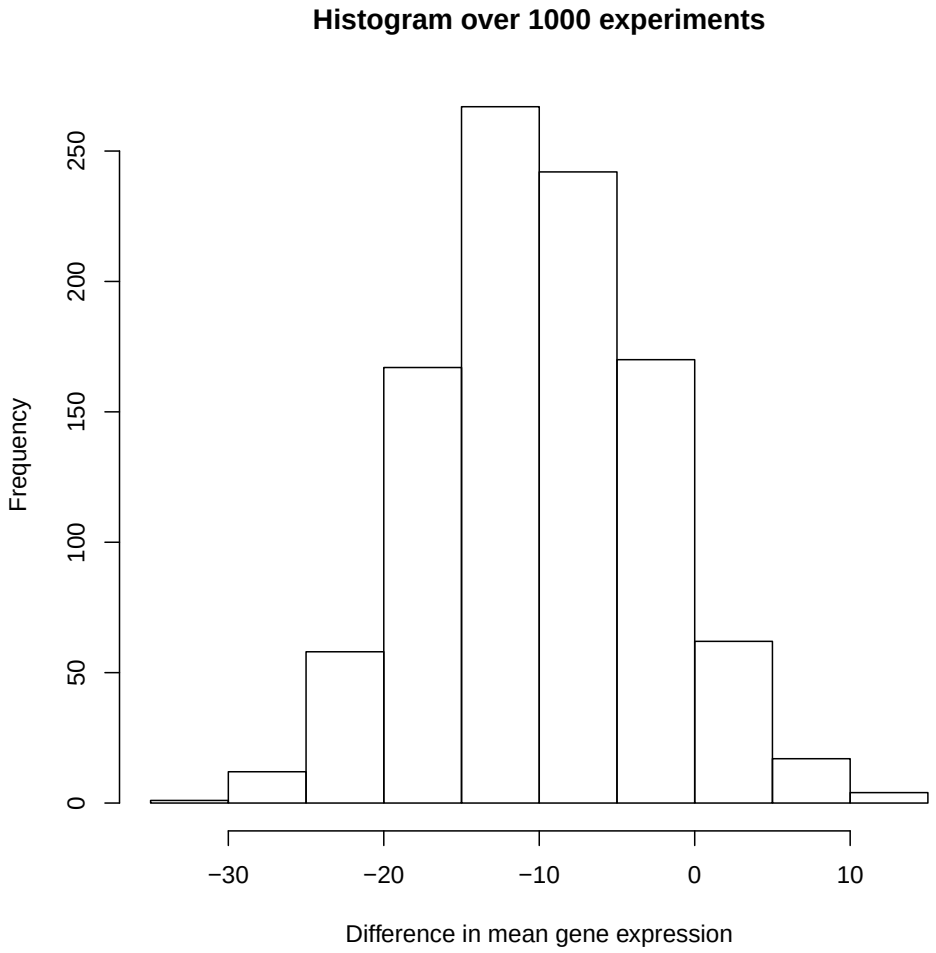

Figure 2:Distribution of differences in mean expression of gene X at E9.5 versus E11.5 over 1000 experiments. Each experiment involved the random sampling of 4 mouse embryos at each of E9.5 and E11.5 developmental time-points and then assaying the expression of gene X over these 8 samples

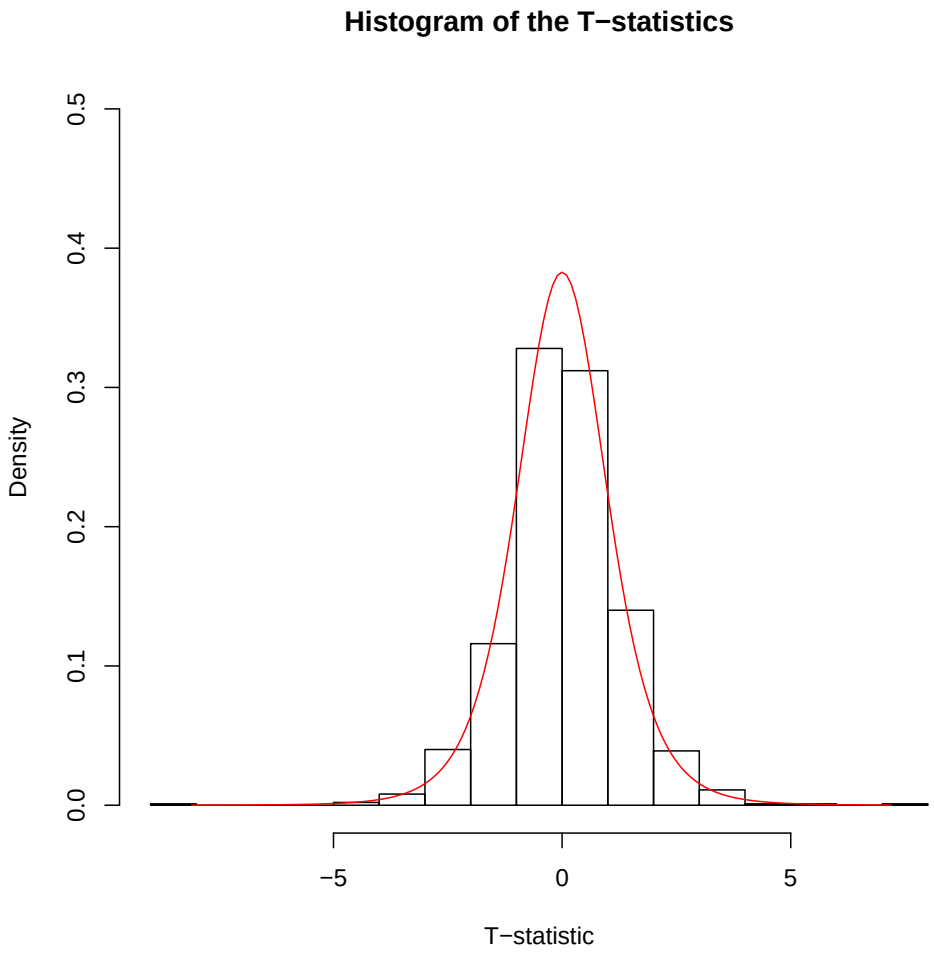

Figure 3:Distribution of t-statistic (a measure of the differences in mean expression at E9.5 versus E9.5) over 1000 experiments. Each experiment involved the random sampling of 8 mouse embryos at E9.5 developmental time-point and then assaying the expression of gene X over these 8 samples. 4 of the these 8 samples are randomly assigned to one group while the rest to the second group. It should be clear that you wouldn't expect a difference, hence this distribution is centered on 0. The solid red line represents theoretical distribution of these t-statistics. Theoretical, i.e., without doing the experiments a 1000 times

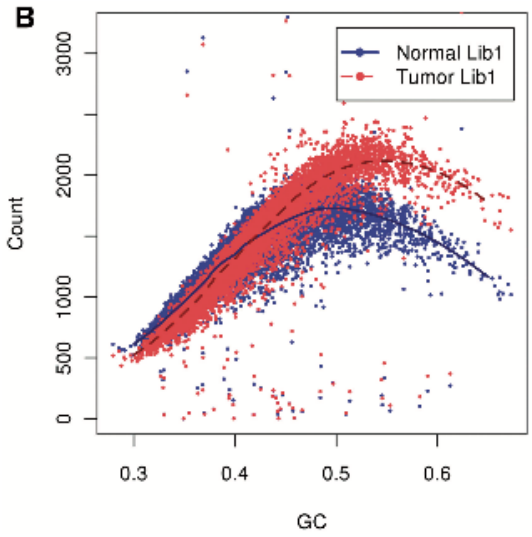

Figure 4:The counts of fragments sequenced using Next-Generation sequencing technology is dependent not only on the sample type (here it is tumor versus normal) but also on the GC content of the fragment. So if you are trying to estimate copy number variation (CNV) between tumor and normal samples then there may be regions you incorrectly identify as having a CNV if you fail to take into account the different relationships between fragment count and GC content in the tumor and normal samples. *This figure is taken from Benjamini and Speed, 2012.*

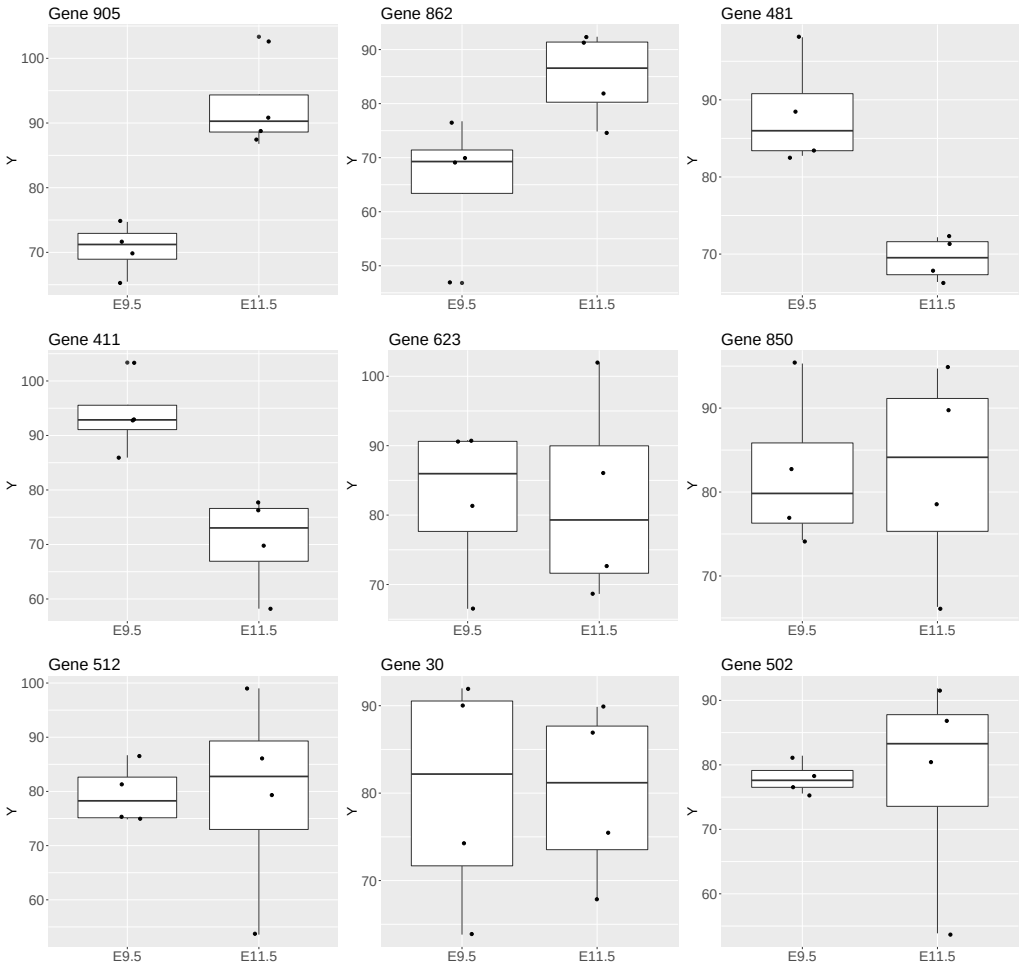

Figure 5:Distribution of observed expression of 9 genes over 4 replicates at two developmental time-points. It appears that the first four genes: 905, 862, 481 and 411 are differentially expressed between the two time-points. However, this is not so because "I played God" - I simulated the expression for each of these 9 genes along with 991 others with the assumption that the mean levels of these genes are not different between the two time-points. So what you see appears like differential expression is actually the result of random variation and not true difference. This points to the need for correcting for multiple testing when your null hypothesis involves simultaneous tests across multiple (> 1) end-points