Lecture 3
Pre-processing serological data

May 21, 2025

Seroanalytics Training
Blantyre, Malawi

Lecture outline

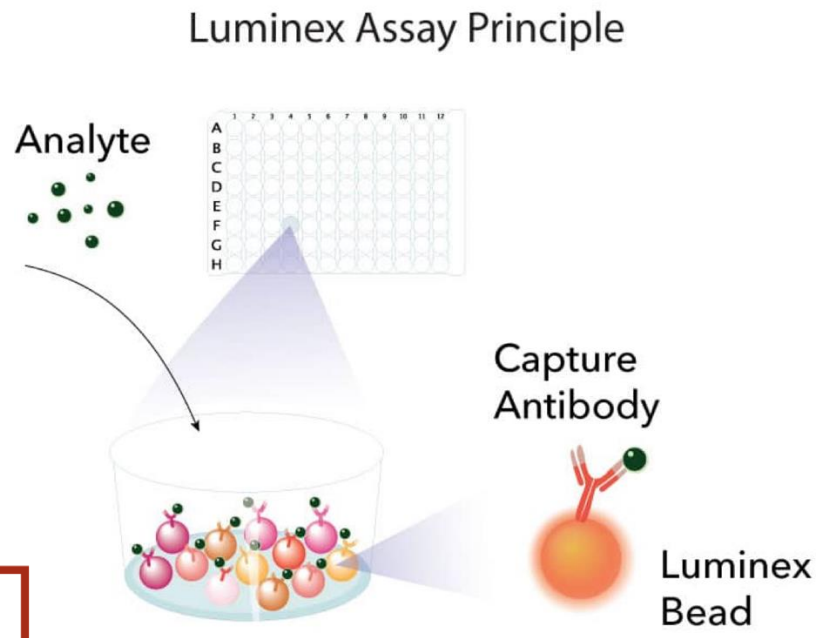
- What is pre-processing and why is it important?
- Can we measure whether pre-processing is effective?
- Pipeline for pre-processing data
  - Filter for quality (bead count)
  - Background correction
  - Transformation
  - Standardization
  - Normalization

What is preprocessing?

- **Pre-processing** is a serries of steps taken to **transform data before formal analysis** and inference (e.g., hypothesis testing) is undertaken.
- Usually, the goals of pre-processing are to **get the data "in shape"** to apply a particular statistical or mathematical test or model.
- Getting data in shape can include several goals:
  - checking distributional assumptions for parametric tests
  - log transformation for visualization
  - **reducing technical variation**

Raw MFIs contain biological and technical variation

Biological variability:
- Differences across samples (natural person-to-person variability)
- Differences across disease condition



Laboratory variability:
- Temperature
- Sample and reagent storage

Machine variability:
- Time since last calibration, maintenance, cleaning
- PMT settings

User variability:
- Pipetting precision

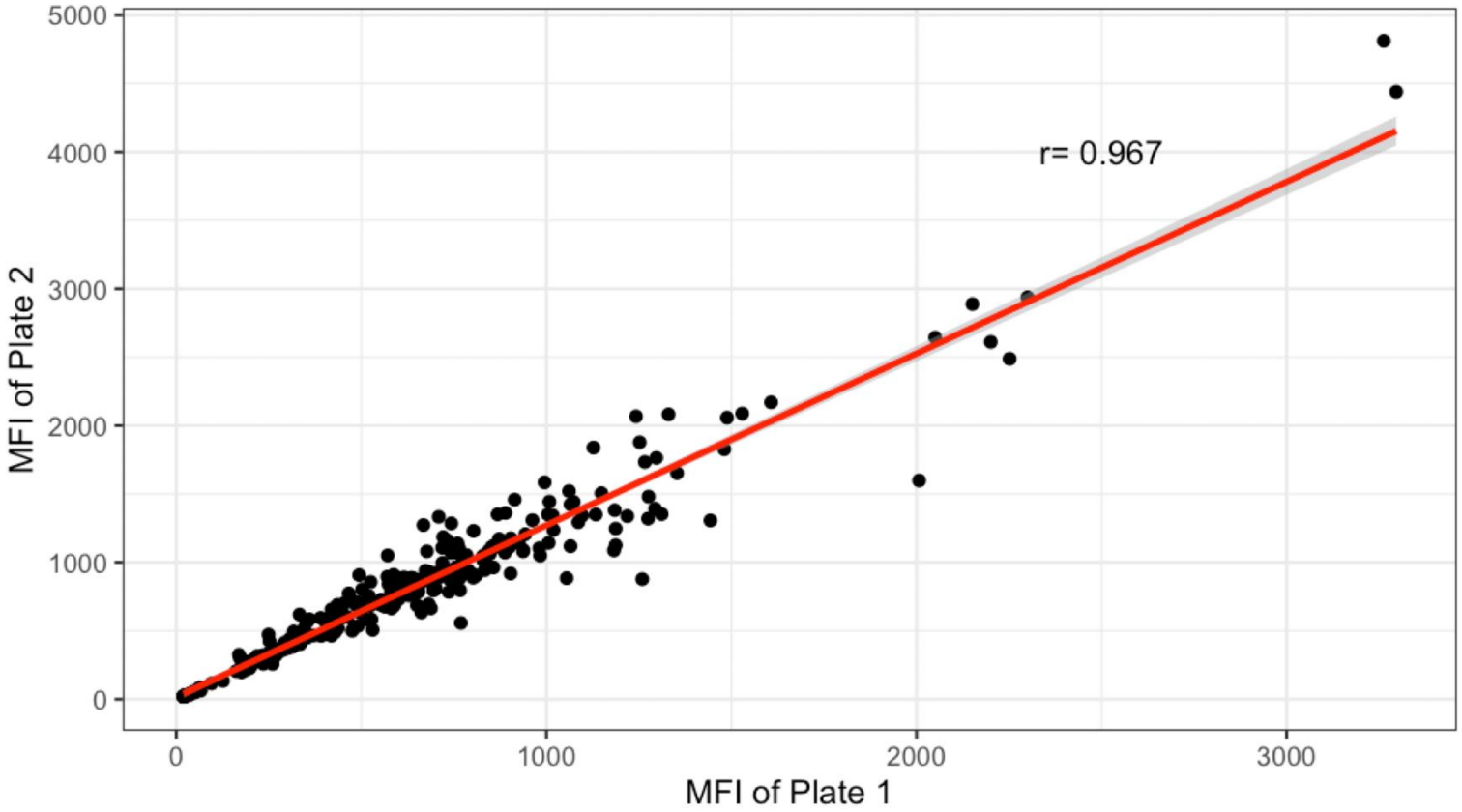
Why is pre-processing important?

- What might happen if we don't perform any pre-processing whatsoever?
- We may **apply a parametric statistical test** (a statistical test that makes distributional assumptions about data) that is **inappropriate**.
- If we leave a substantial amount of technical variation in our data, we may **attribute** observed differences to **biological mechanisms** when they are in fact the result of **technical differences**.

How can we assess whether pre-processing is effective?

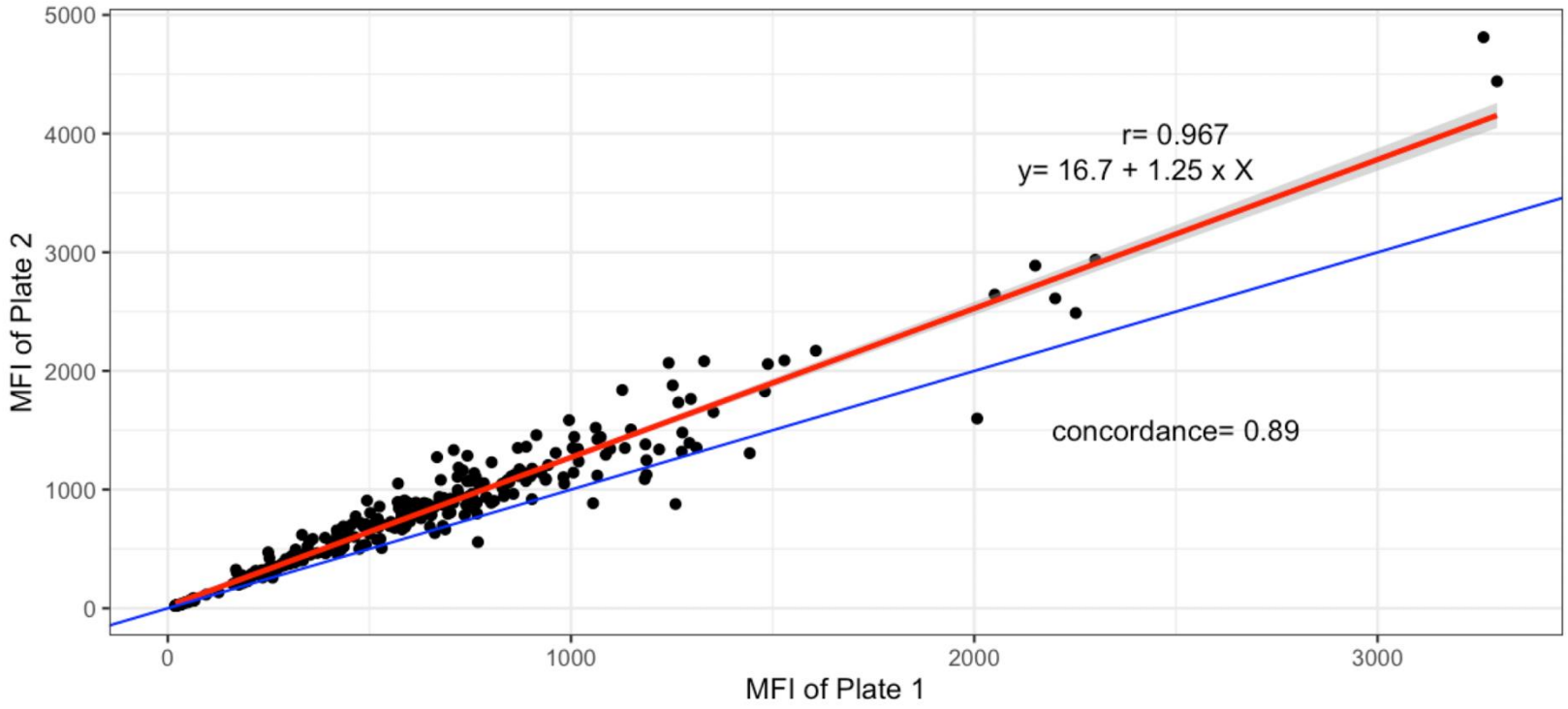
- We will focus on the goal of **removing technical variation.**
- We will use **technical replicates** to accomplish this goal, that is, **samples that we re-measure** at various points during the assay.
- The assumption we rely on is that any differences we measure across these samples **cannot be biological** in nature, and therefore **must be** the result of **technical variation.**

Running same samples on two plates



What happens if we just measure **correlation**?
What would you conclude about **technical differences** between these two plates?

Running same samples on two plates

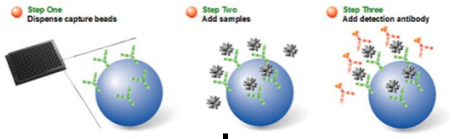


Now, what happens if we measure **correlation specifically around the line y=x**?
What might we conclude about these two plates?

Evaluating whether pre-processing was effective

- By measuring **concordance** across **technical replicates** throughout an assay (e.g., between plates, batches, or other assay units), we can assess the level of **technical variation**.
- The goal of removing technical variation is often **not accomplished** with **one single transformation** or step; **multiple steps** may be necessary.
- The **ordering** of these steps is also important.
- A particular **series and ordering** of steps that pre-process data is referred to as a **pre-processing pipeline**.

An example pipeline for pre-processing data



1. Read in CSV, Standard Curves, Controls, Background

2. Filter based on bead count threshold

3. Remove background

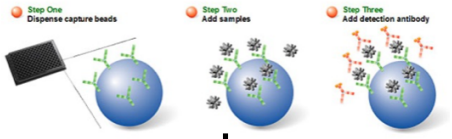
4. Transform (Natural Log)

5. Fit standard curves and estimate log concentrations

6. Normalize

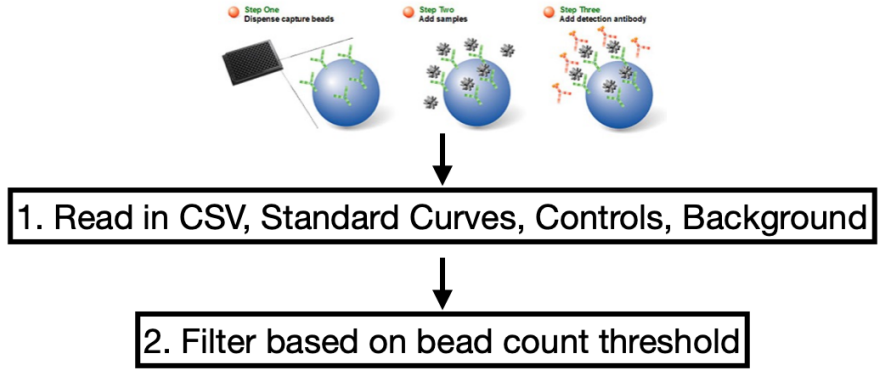
An example pipeline for pre-processing data

This step can also ensure data are in ‘tidy’ format.



1. Read in CSV, Standard Curves, Controls, Background

An example pipeline for pre-processing data



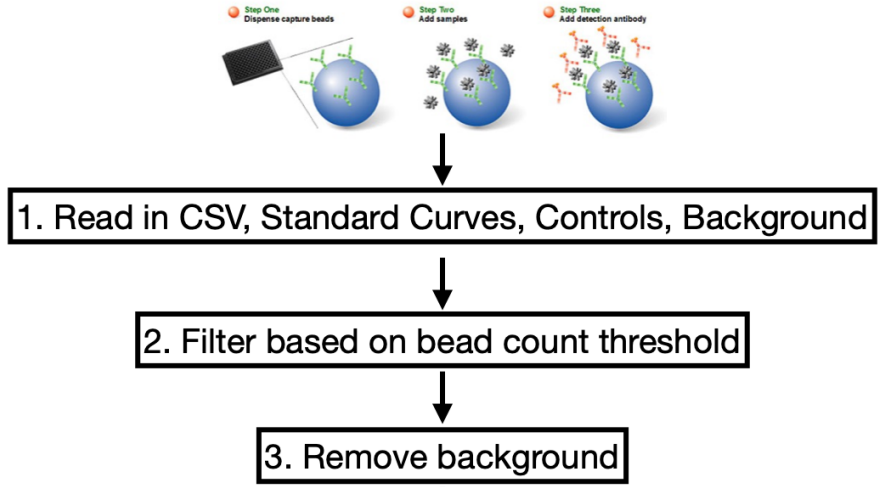
Bead counts

- In bead-based assays, **antigens** of interest are **bound to beads** that have a **specific identification.**
- Since these are multiplexed assays, we may be interested in measuring antibody responses to **more than one type of antigen**.
- Each **type of antigen** is bound to a **set of beads** with **a unique identifier.**
- This way, the **number of beads** read for **each antigen** can be computed, and the **fluorescence for each antigen** is taken as the **median** of the measured fluorescence **for all beads with the identifier** corresponding to that specific antigen.

Bead counts

- When a **very small number** of beads for a particular antigen (one bead ID) are read by the machine, this is **usually an issue**.
- It can be the result of **technical difficulties** with the assay or the fluidics of the reading or scanning machine.
- In these cases, the **median fluorescence** measurement for this particular antigen **can be unreliable**.
- Therefore, it is common to set a threshold for a **minimum number of beads** required to proceed with analysis for a particular antigen in a specific sample (e.g., 30 or 50 beads are common thresholds).

An example pipeline for pre-processing data



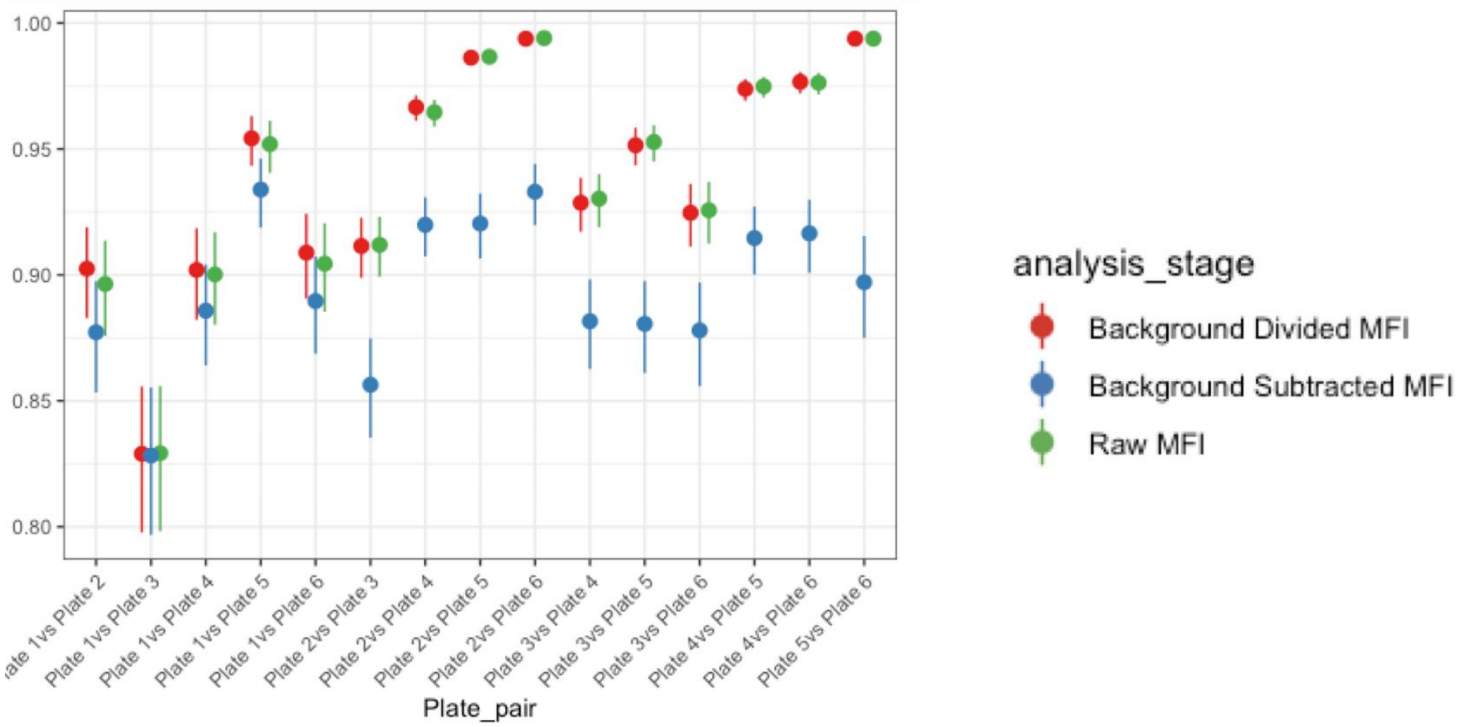
What is background?

- Background is signal that is not the result of a specific binding event between an antigen and an antibody.
- It is often the result of autofluorescence, that is fluorescence resulting from the reagents used for the assay.
- Therefore, background is typically measured with ‘blank’ wells, that is wells that contain all reagents for the assay but no sample.
- These are usually placed at least once on every plate, but can be present more than once.

Background correction

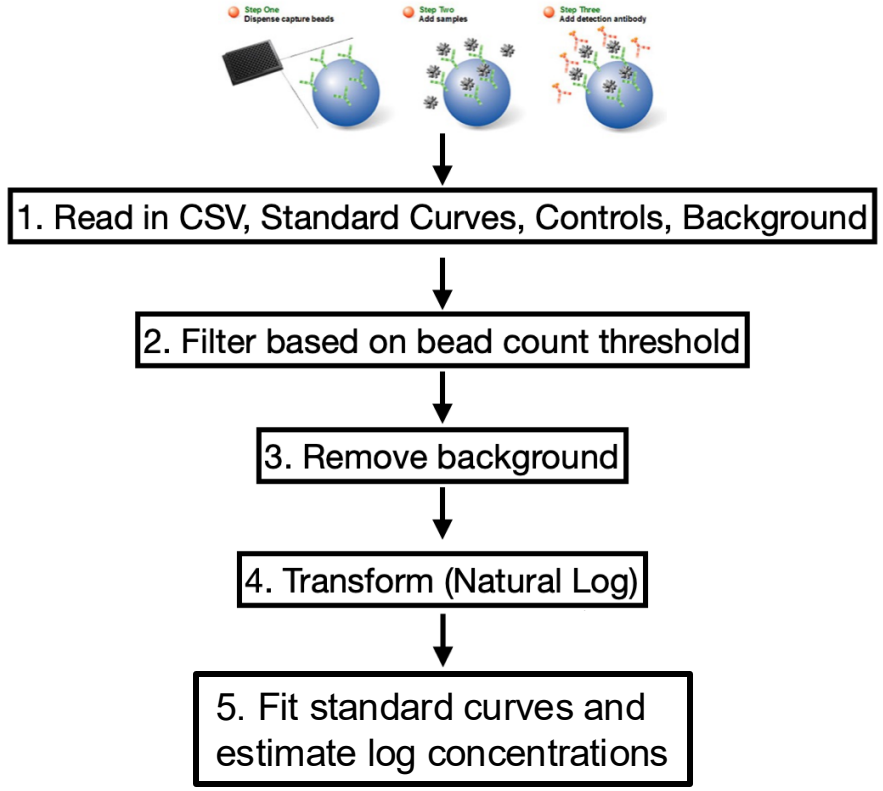
- Techniques to remove background include:
  - **Subtraction**: MFI of sample – MFI blank
  - **Division**: MFI of sample / MFI blank
- This is typically done on a plate-by-plate basis; therefore, if there is more than one blank well per plate, “MFI blank” could be the median or mean of all blank wells on that plate.
- An important factor to consider is that **if the MFI of a sample is lower or equal to the MFI of blank well(s),** the subtraction may produce a **value less than or equal to 0**.

Example results of background correction



MBA for malaria study (6 x 384 well plates, 19 P. falciparum antigens)

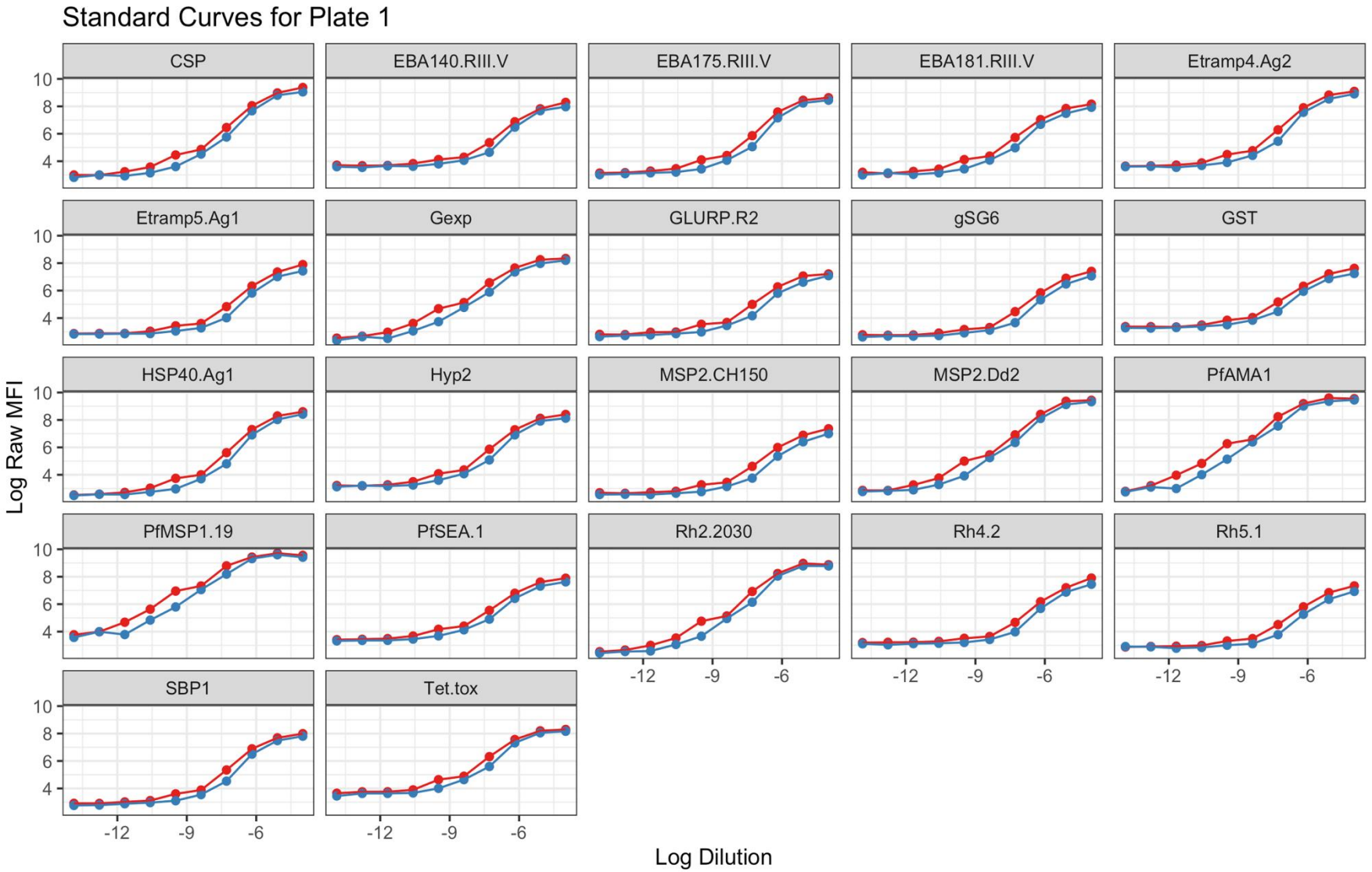
An example pipeline for pre-processing data



Standardization

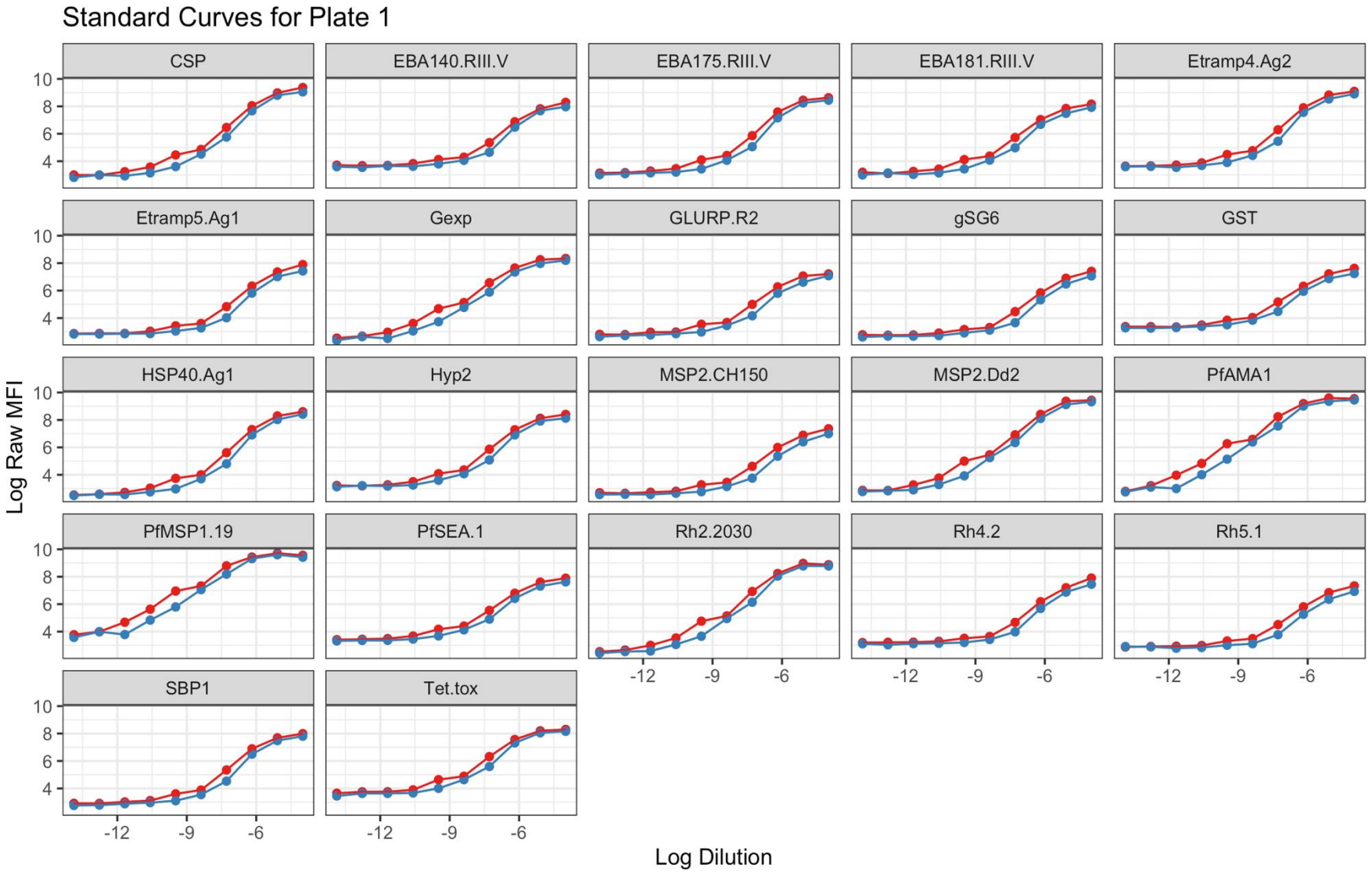
- Standardization changes MFI measurements to other units of concentration that can be useful for certain antigens.
- Examples of concentration units include:
  - **Relative Antibody Units (RAU)**
  - **International Units (IU):** can correspond to WHO guidelines for protective immunity
- This process is accomplished by having **serial dilutions of a sample** with **known concentrations** (measured as a dilution factor, RAU, IU, etc). This serial dilution is then run on one or more plates and a relationship between concentration and MFI is established.

Example of a standard curve:



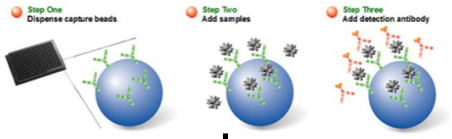
- This “S” shaped curve is the ideal shape between log MFI and log concentration (in this case, measured as a dilution factor).
- It is called a **logistic curve** and typically only exists when you observe both **MFI and concentration** on the **log** scales.

Example of a standard curve:



- We can estimate the **equation** that describes each logistic curve.
- Then, using that equation, we can **transform** a specific sample's **MFI** into an estimate of **concentration**.
- This can be done on a **plate-by-plate basis** if standard curves are available on all plates.
- Or, it can be done **once for an entire study** if there is only one standard curve available.

An example pipeline for pre-processing data



1. Read in CSV, Standard Curves, Controls, Background

2. Filter based on bead count threshold

3. Remove background

4. Transform (Natural Log)

5. Fit standard curves and estimate log concentrations

6. Normalize

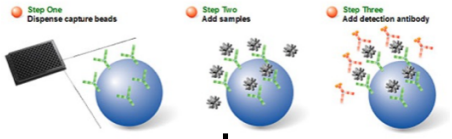
Normalization

- Oftentimes, when we run an experiment in the lab it is not possible to run all the samples from a study all at once.
- They are often broken down into **smaller units** (e.g., plates) and then even these **units can be run at different times** (e.g., in different batches).
- Normalization **seeks to remove differences between units or batches**.
- We will work with a simple method (there are many!)
- We will use technical replicates that are measured repeatedly in each batch to normalize.

Normalization method

- We will work on the assumption that **differences across these technical replicates** are strictly due to **technical variation**.
- Therefore, we want to **remove these differences**.
- We will estimate the **mean difference** due to technical variation between each plate by fitting a **linear model to the technical replicate data** which is on the log concentration scale (recall, we just fit our standard curve!).
- Then we will **subtract this estimated mean difference** from all log concentrations, that is **cohort samples**.
- These normalized values will then be used for downstream analysis.

An example pipeline for pre-processing data



1. Read in CSV, Standard Curves, Controls, Background

2. Filter based on bead count threshold

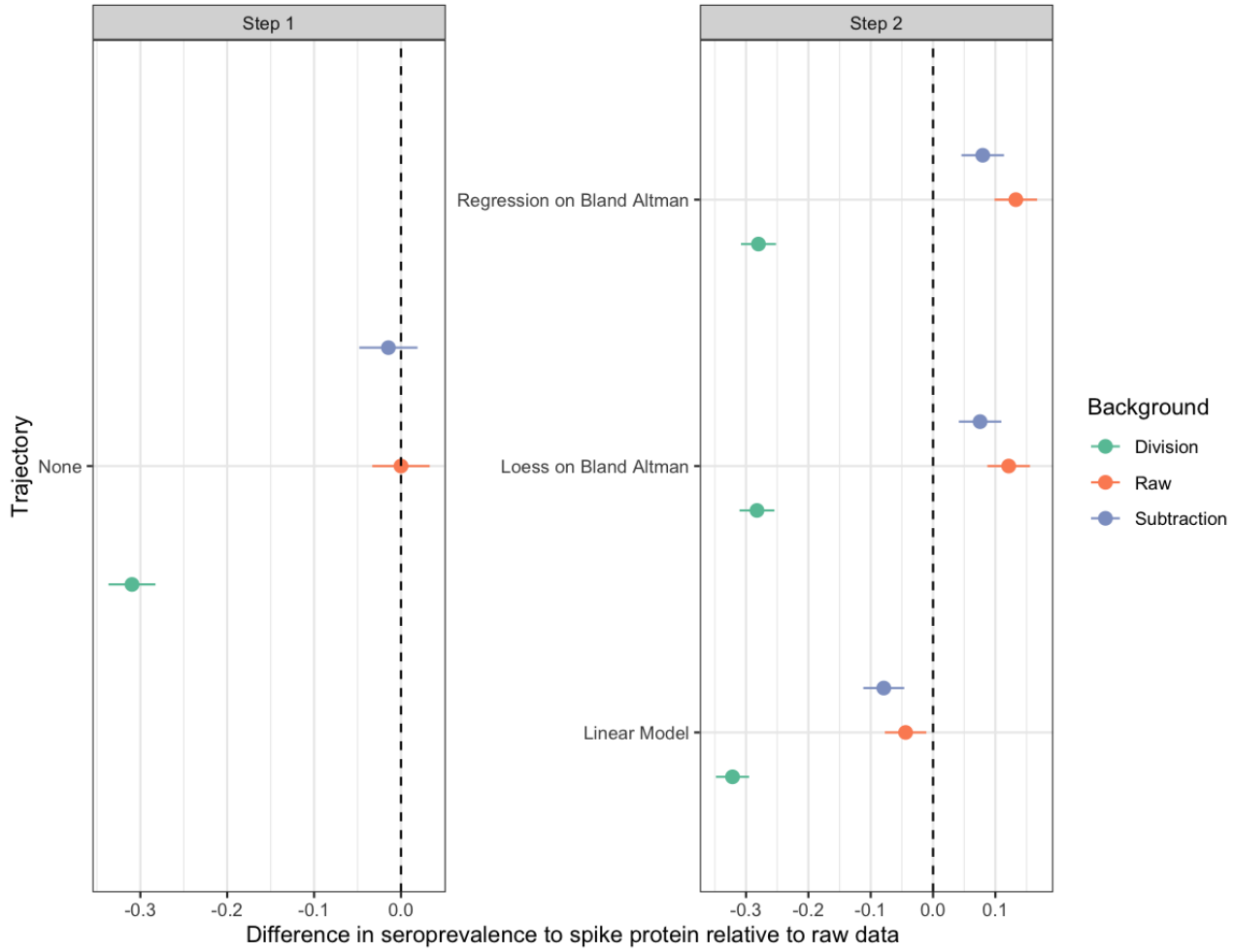
3. Remove background

4. Transform (Natural Log)

5. Fit standard curves and estimate log concentrations

6. Normalize

A quick example of the impact of pre-processing.



Conclusion

- Pre-processing can directly **remove** some **technical variation** and ensure our data are “in shape” to perform our **analysis** and **get accurate inference**.
- The ”best” combination of pre-processing steps can be evaluated by **measuring concordance across technical replicates** and may vary across data.
  - This means that **technical replicates are important** when you are considering assay design!
- Pre-processing **makes a difference** in downstream inference.