RANDOMIZE

Introduction:

High-throughput DNA methylation arrays are susceptible to bias facilitated by batch effects and other technical noise that can alter DNA methylation level estimates. RANDOMIZE is a user-friendly web application that provides an interactive and flexible graphical user interface (GUI) to randomize relevant metadata. Using this tool will minimize chip and position mediated batch effects in microarray studies for an increased validity in inferences from your methylation data. The tool is very helpful for biologist to perform randomization of test samples and insert controls in the data. The tool is available online at https://coph-usf.shinyapps.io/RANDOMIZE/ and it is free to use.

Tutorial:

For the assessment of functions and illustration of tool utility, we have used sample data with 750 samples. The sample data is available with the tool.

Go to the **Analysis** tab to start randomization. On the right side is the **Randomization** panel, shown below in Figure 1, where you can browse your computer in order to locate your metadata file and upload the metadata file in a CSV file format. Successful uploads will be indicated as "Upload complete". It is important that your metadata file include columns labeled as **ParticipantID** and **SampleID**.

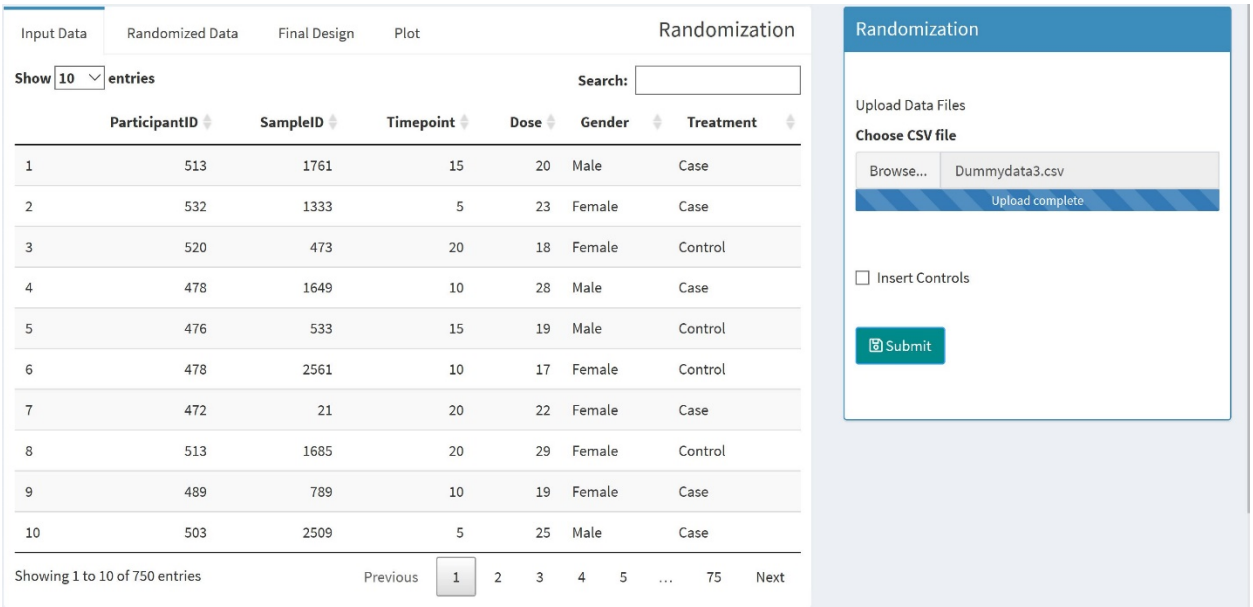

Figure 1: shows input data

Your data should show up in the **Input Data** tab on the top left. In the Randomization panel, you can also check the box "**Insert controls**" to manually add known controls to your analysis. Controls can then be added on individual chips. No controls are inserted by default.

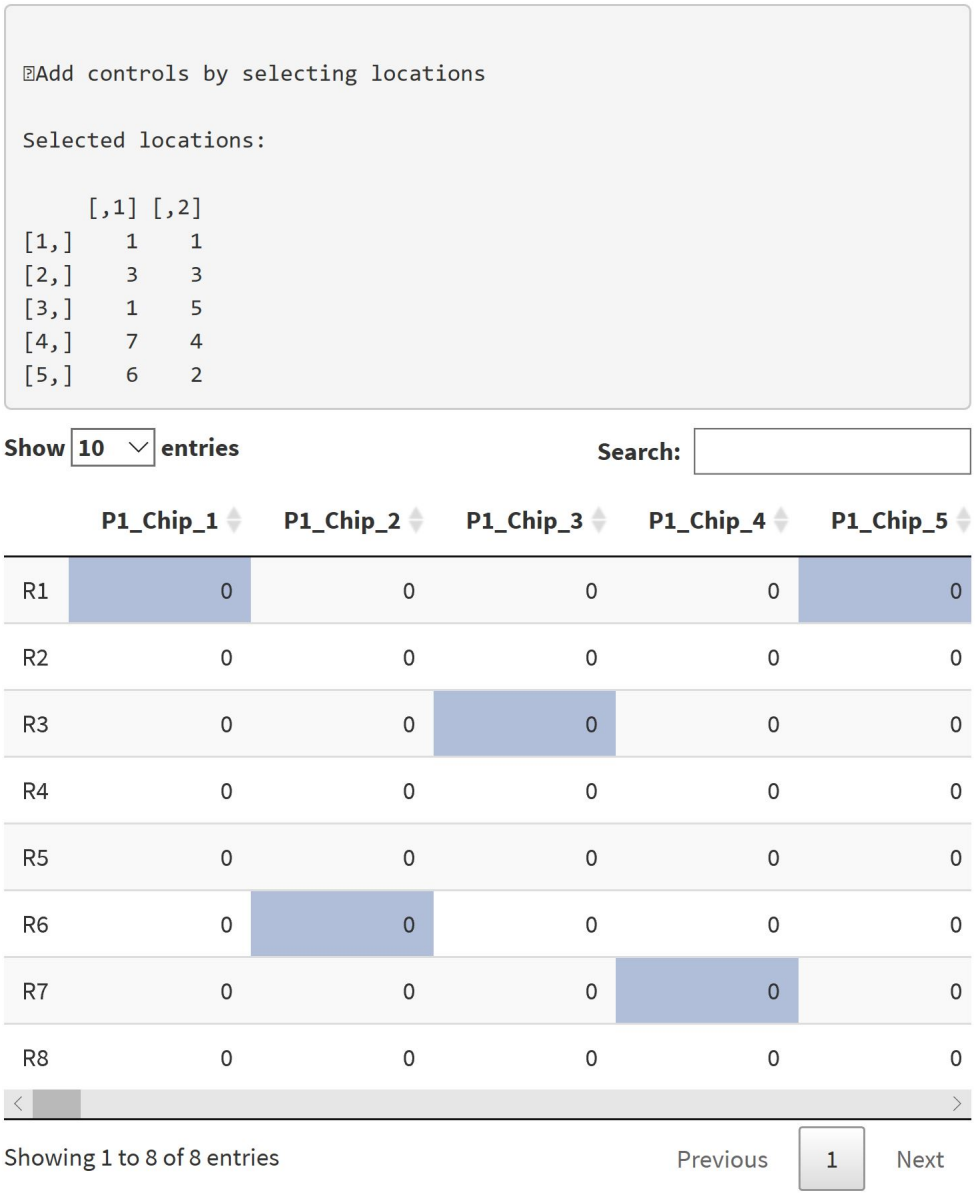

Figure 2: shows places for controls.

By clicking on the columns, you are able to select the columns you would like to randomize (shown below in Figure 3). If you hit "Submit" on the randomization panel without selecting any columns, you will receive an error message: "Please select columns for randomization by clicking on desired column(s)".

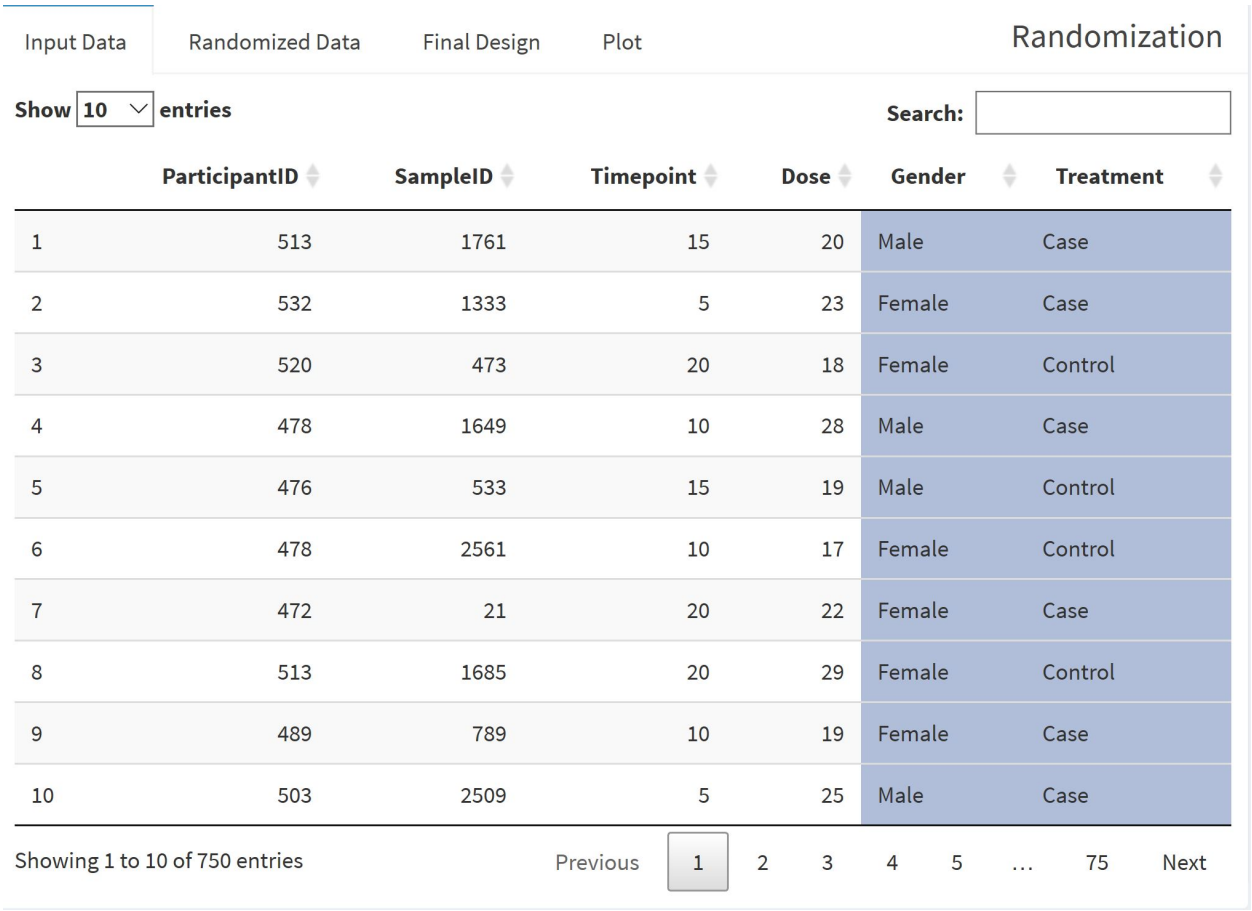

Figure 3: Selected columns of interest for randomization are highlighted in blue.

After selecting the columns of interest for randomization click on the **Submit Button** located on the bottom of the Randomization Panel to submit the job for processing. Once the job is processed, in the **Randomized Data** tab (next to the **Input Data** tab) you can take a look at your randomized data of your selected items, as shown below. Your previously selected controls are excluded from the randomization and still in the location you have selected beforehand. The controls are shown as zeros, as in Figure 4.

Figure 4: shows randomized data

The **Final Design** tab adjacent to the **Randomized data** tab shows you the final design of your randomized data. The **Display Final Data** tab lets you view your final design, one plate at a time. The design of the first plate is available to view by default.

The **Plot** tab eventually shows you your plotted data. You should select the columns of interest by clicking on them before you move on to the **Plot** tab. In the **Plots** selection on the left you are able to choose between various plots. The **Plot labels** option lets you select a title for your plot and label the x- and y-axis.

In this tutorial, we illustrate the goodness of randomization using the sample dataset and sunflower plot. The sunflower plot is used to display bivariate distribution. Each petal on the sunflower plot represents an observation(sample). The **ParticipantID** column in our sample dataset denotes the participant ids; each participant has one or more samples in the range of 1-23. There are 112 unique participants in the dataset. For 750 samples, eight samples on one chip, we need 94 chips in total. An ideal randomization would be that no two or more samples from the same participant are on the same chip; however, the number of chips is less than the number of participants, so it is evident that some samples from the same participant will be on the same

chip. The black dots in Figure 5 denote unique samples. If two samples from the same participant are on the same chip, a petal, as shown in red, is added on the black dot. For two duplicates, two petals are added, and so on. The plot indicates proper randomization of the data — for example, for the participant which has 23 samples, all the samples are sent to different chips. Only some chips have two samples from the same participant id. Similarly, the randomization of participant ids on plates is shown in Figure 6.

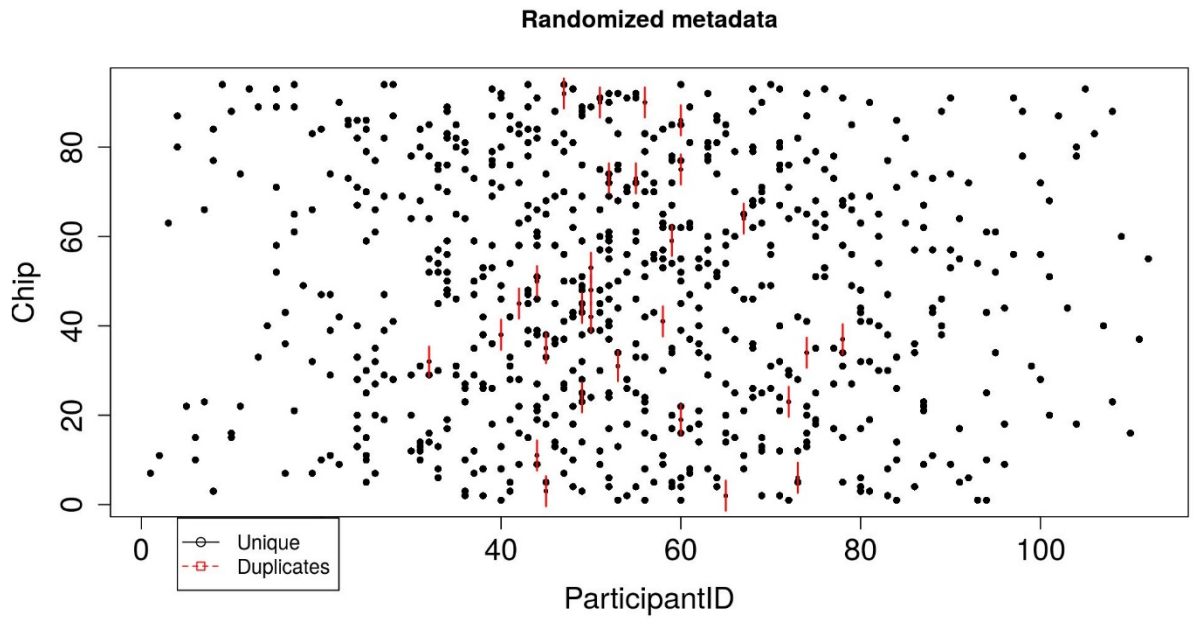

Figure 5: Sunflower plot showing the metadata distribution (here, participantID) across Chips.

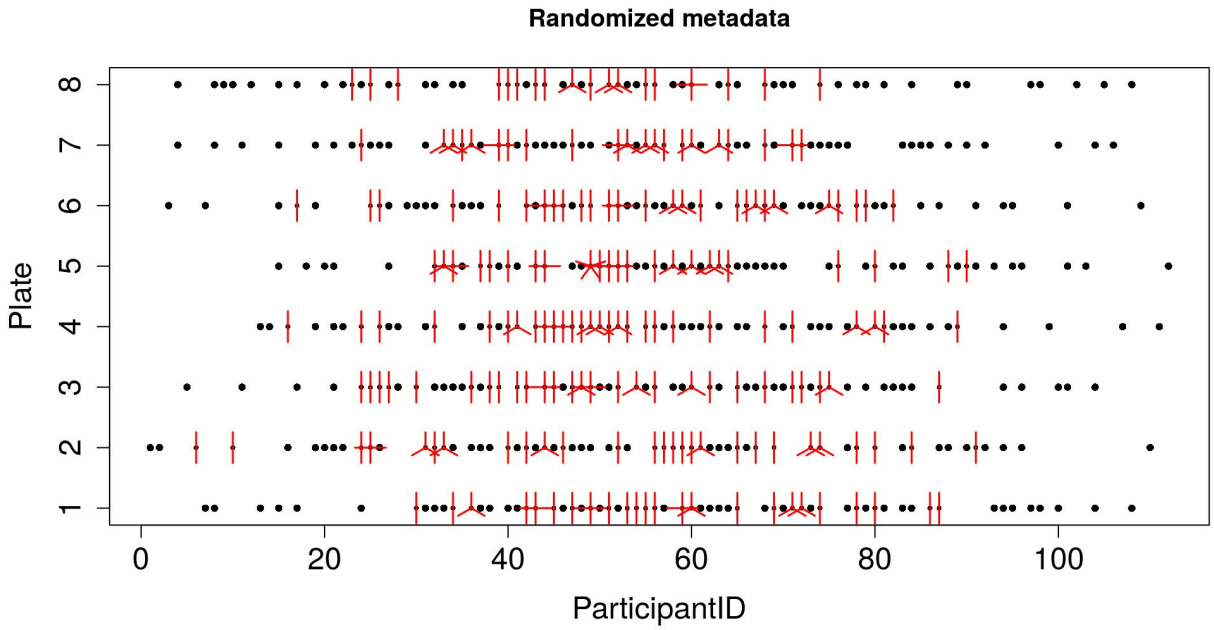

Figure 6: Sunflower plot showing the metadata distribution (here, participantID) across Plates.

Download

The **Download** tab lets you download your randomized data. In the **Randomized Data** tab, you are able to download your **Randomized Data** and **Final Design** in a CSV file format while an individually downloaded **Plot** will be in a PNG file format.