

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**.

The screenshot displays the RANDOMIZE web application interface. The main panel is titled 'Input Data' and shows a table of 10 entries. The table has columns for ParticipantID, SampleID, Timepoint, Dose, Gender, and Treatment. The data is as follows:

	ParticipantID	SampleID	Timepoint	Dose	Gender	Treatment
1	513	1761	15	20	Male	Case
2	532	1333	5	23	Female	Case
3	520	473	20	18	Female	Control
4	478	1649	10	28	Male	Case
5	476	533	15	19	Male	Control
6	478	2561	10	17	Female	Control
7	472	21	20	22	Female	Case
8	513	1685	20	29	Female	Control
9	489	789	10	19	Female	Case
10	503	2509	5	25	Male	Case

Below the table, it says 'Showing 1 to 10 of 750 entries' and has pagination controls for 'Previous', '1', '2', '3', '4', '5', '...', '75', and 'Next'.

On the right side, there is a 'Randomization' panel. It has a title 'Randomization' and a section 'Upload Data Files'. Under this section, it says 'Choose CSV file' and has a 'Browse...' button. A file named 'Dummydata3.csv' is shown with an 'Upload complete' status. There is a checkbox for 'Insert Controls' and a 'Submit' button.

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.

☒ Add controls by selecting locations

Selected locations:

	[,1]	[,2]
[1,]	1	1
[2,]	3	3
[3,]	1	5
[4,]	7	4
[5,]	6	2

Show entries
 Search:

	P1_Chip_1	P1_Chip_2	P1_Chip_3	P1_Chip_4	P1_Chip_5
R1	0	0	0	0	0
R2	0	0	0	0	0
R3	0	0	0	0	0
R4	0	0	0	0	0
R5	0	0	0	0	0
R6	0	0	0	0	0
R7	0	0	0	0	0
R8	0	0	0	0	0

Showing 1 to 8 of 8 entries
 Previous

 Next

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)”.

<div> <div>Input Data</div> <div>Randomized Data</div> <div>Final Design</div> <div>Plot</div> <div>Randomization</div> </div>						
Show <div>10</div> entries				Search: <input type="text"/>		
	ParticipantID	SampleID	Timepoint	Dose	Gender	Treatment
1	513	1761	15	20	Male	Case
2	532	1333	5	23	Female	Case
3	520	473	20	18	Female	Control
4	478	1649	10	28	Male	Case
5	476	533	15	19	Male	Control
6	478	2561	10	17	Female	Control
7	472	21	20	22	Female	Case
8	513	1685	20	29	Female	Control
9	489	789	10	19	Female	Case
10	503	2509	5	25	Male	Case
<div> <div>Showing 1 to 10 of 750 entries</div> <div> <div>Previous</div> <div>1</div> <div>2</div> <div>3</div> <div>4</div> <div>5</div> <div>...</div> <div>75</div> <div>Next</div> </div> </div>						

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.

Input Data

Randomized Data

Final Design

Plot

Randomization

Show

10

▼

entries

Search:

	ParticipantID	SampleID	Timepoint	Dose	Gender	Treatment	plate	chip
1	0	0	0	0	0	0	plate_1	chip_1
2	515	2601	10	22	Male	Case	plate_1	chip_1
3	480	2701	10	19	Female	Control	plate_1	chip_1
4	521	1977	5	20	Female	Control	plate_1	chip_1
5	505	345	20	19	Female	Control	plate_1	chip_1
6	480	2337	15	21	Female	Case	plate_1	chip_1
7	497	2165	15	14	Female	Case	plate_1	chip_1
8	484	1157	20	11	Male	Control	plate_1	chip_1
33	492	2709	10	17	Female	Control	plate_1	chip_2
34	495	369	10	19	Male	Case	plate_1	chip_2

Showing 1 to 10 of 752 entries

Previous

1

2

3

4

5

...

76

Next

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.

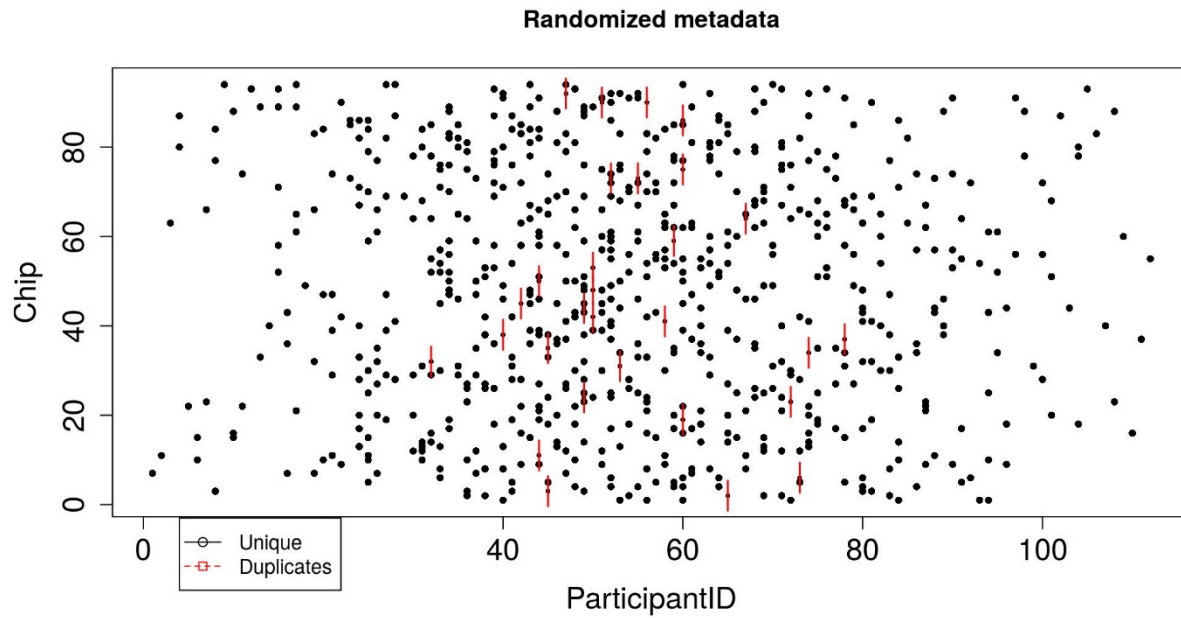


Figure 5: Sunflower plot showing the metadata distribution (here, participantID) across Chips.

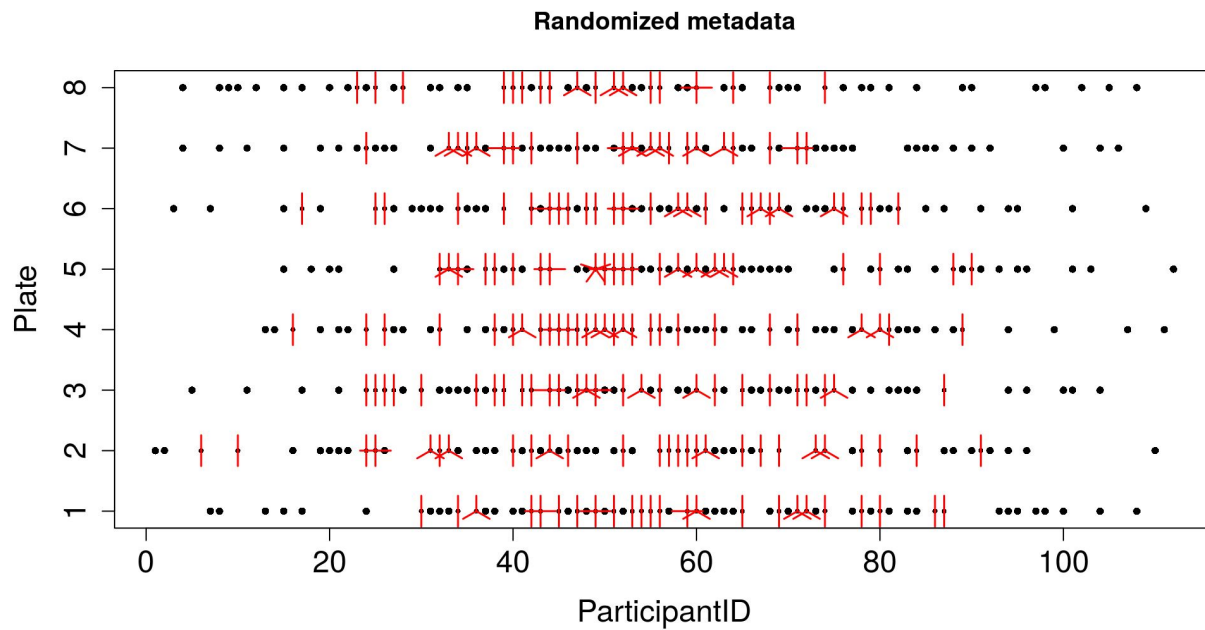


Figure 6: Sunflower plot showing the metadata distribution (here, participantID) across Plates.

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.