Toward enhanced reproducibility in biological experiments : mix and pick !

Elise JACQUEMET [1], Juliette MEYER [1], Thomas OBADIA [1], Hugo VARET [1], Pascal CAMPAGNE [1], Stevenn VOLANT [1]

[1] Institut Pasteur, Bioinformatics and Biostatistics Hub, Université Paris Cité, 75015 Paris, France

Why controlling randomisation ?

Assigning individuals to experimental groups at random can introduce serious biases and confounding factors into an experiment whenever important covariates are not properly accounted for. Indeed, the outcome of interest may be strongly affected by hidden effects (e.g., age, genetic background, technical factors, etc.) due to an uneven distribution of covariates among groups. Such biases may obscure the genuine experimental effects, reduce statistical resolution, and increase the rate of false positives.

Population
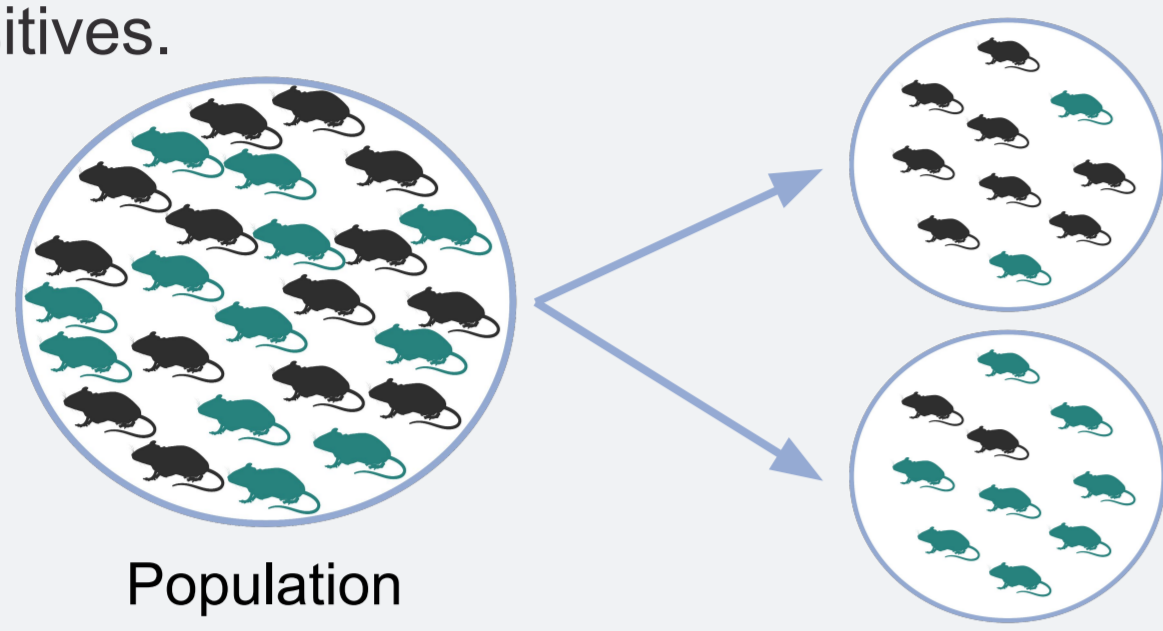
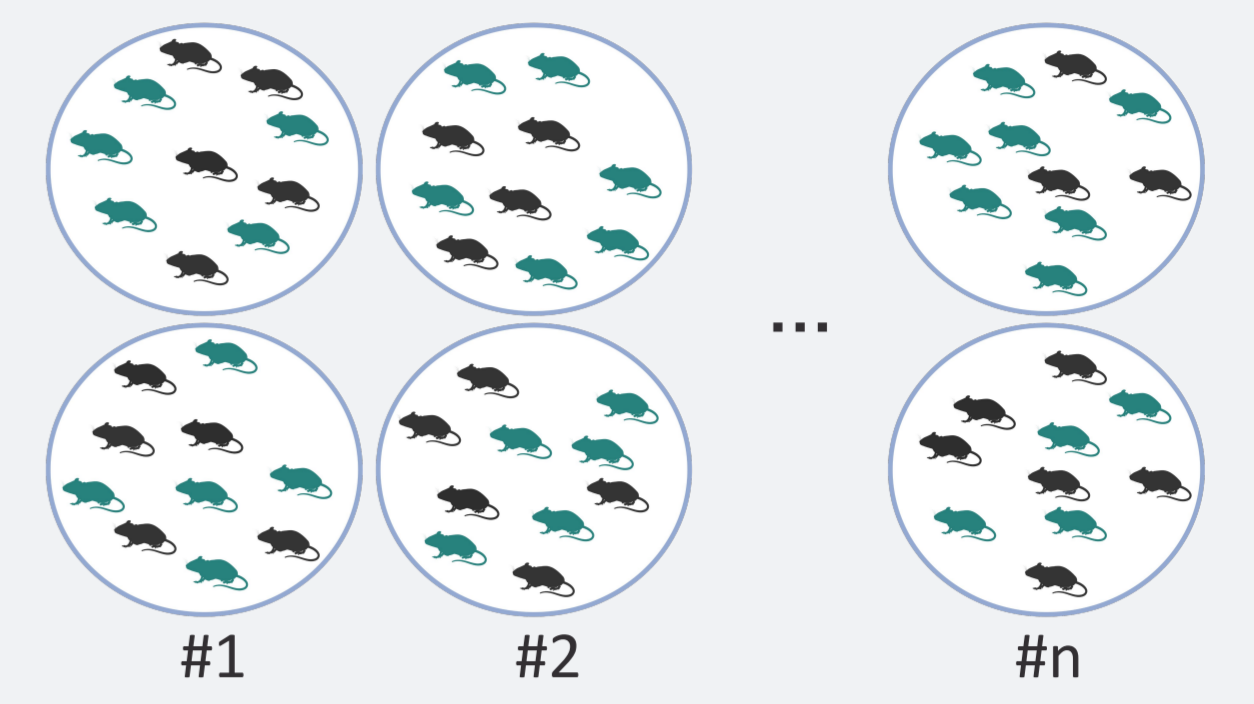*Example of an unlucky randomisation*

Our approach

We propose a user-friendly web interface to assist randomisation processes. It aims at minimizing bias while enhancing both reliability and interpretability of statistical analysis, by controlling covariates during randomization.

mixnpick.pasteur.fr

Multiple possible assignments (ranked by score)

#1   #2   ...   #n

Loading data

3-steps loading process

1. Load a tabulated file (e.g., excel file)
2. Detect variable type (qualitative or quantitative)
3. Data are checked. Warnings and/or error are displayed if there are any

Select a file

Load a file from your computer. Supported format are .xls or . xlsx files. If you would like another format to be supported, feel free to submit a request by mail.

Browse...   Load your data

Loaded data

ID	cage	born	sex	batch	weight	CD4
#2023_1	A	16	F	a	20.2	94.1
#2023_2	A	3	M	a	21	94.9
#2023_3	A	13	M	a	19	90.7
#2023_4	A	8	F	a	21.8	83.3

Description

You can edit the lists by clicking on pencil symbols.

Quantitatives variables ✏   Qualitatives variables ✏   Individuals identifier ✏
• born   • cage   • ID
• weight   • sex
• CD4   • batch
• ht_cell

No error nor warning has been detected

VALIDATE

Input data

Rows represent individuals. Columns correspond to various features.

The table can contain supernumerary individuals as compared to the needs of the experiment (i.e., some will be assigned to an outgroup cluster).

Selecting variables

Users may tune variable types manually (qualitative or quantitative) with a simple drag-and-drop.
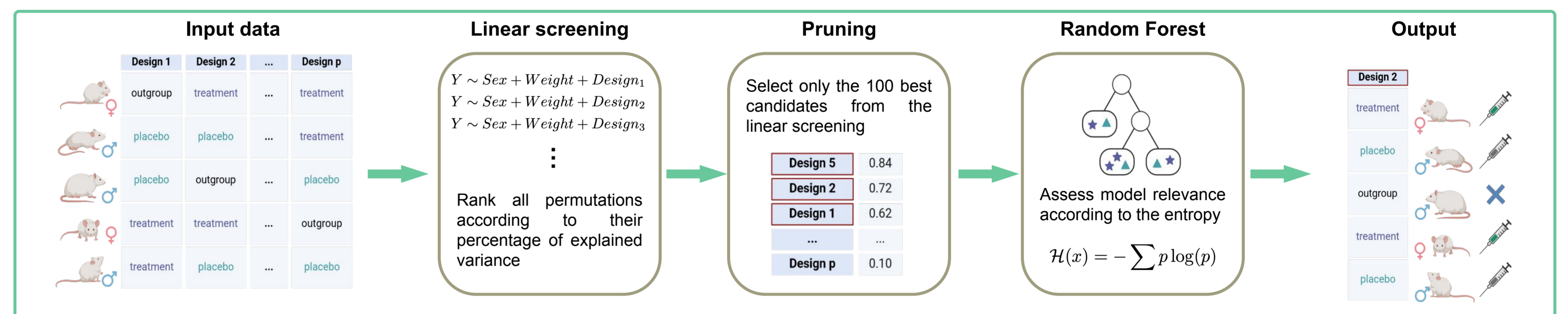
Checking

Summary of the input variables ; notices may appear when any issues are detected.
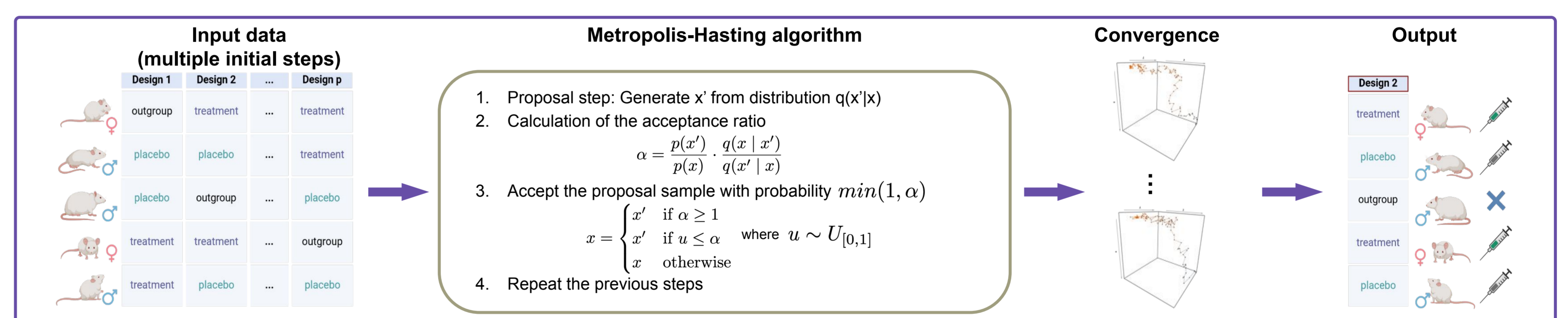
Simulation step

Two algorithms available

Permutation: This method of assigning individuals to groups combines a regular linear model and a random forest classifier. It selects assignments among a set of random permutations (i.e., those with the lower propensity scores) white accounting for both linear and non-linear effects.

Integration: This algorithm is based on Metropolis-Hastings approach. It refines an initial random permutation by iteratively modifying the distribution of individuals so that homogeneity of groups is improved.
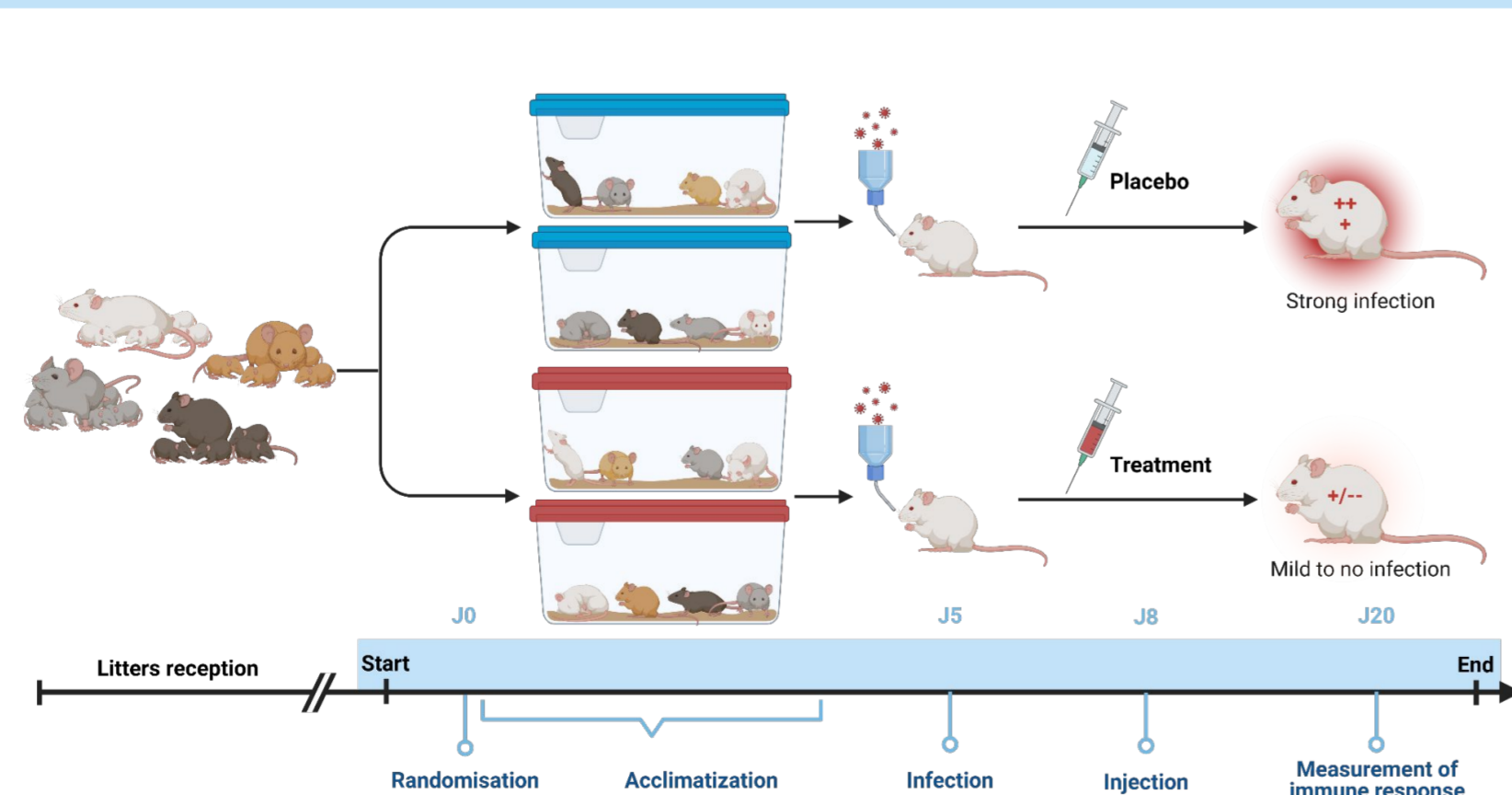
Permutation

Input data			
Design 1	Design 2	...	Design p
outgroup	treatment	...	treatment
placebo	placebo	...	treatment
placebo	outgroup	...	placebo
treatment	treatment	...	outgroup
treatment	placebo	...	placebo

Linear screening

$Y \sim Sex + Weight + Design_1$
$Y \sim Sex + Weight + Design_2$
$Y \sim Sex + Weight + Design_3$
$\vdots$

Rank all permutations according to their percentage of explained variance

Pruning

Select only the 100 best candidates from the linear screening

Design 5	0.84
Design 2	0.72
Design 1	0.62
...	...
Design p	0.10

Random Forest

Assess model relevance according to the entropy

$\mathcal{H}(x) = -\sum p \log(p)$

Output

| Design 2 |
| treatment |
| placebo |
| outgroup |
| treatment |
| placebo |

Integration

Input data (multiple initial steps)		
Design 1	Design 2	Design p
outgroup	treatment	treatment
placebo	placebo	treatment
placebo	outgroup	placebo
treatment	treatment	outgroup
treatment	placebo	placebo

Metropolis-Hasting algorithm

1. Proposal step: Generate x' from distribution q(x'|x)
2. Calculation of the acceptance ratio
$$\alpha = \frac{p(x')}{p(x)} \cdot \frac{q(x \mid x')}{q(x' \mid x)}$$
3. Accept the proposal sample with probability $min(1, \alpha)$
$$x = \begin{cases} x' & \text{if } \alpha \geq 1 \\ x' & \text{if } u \leq \alpha \\ x & \text{otherwise} \end{cases} \text{ where } u \sim U_{[0,1]}$$
4. Repeat the previous steps

Convergence

Output

| Design 2 |
| treatment |
| placebo |
| outgroup |
| treatment |
| placebo |

Outcomes

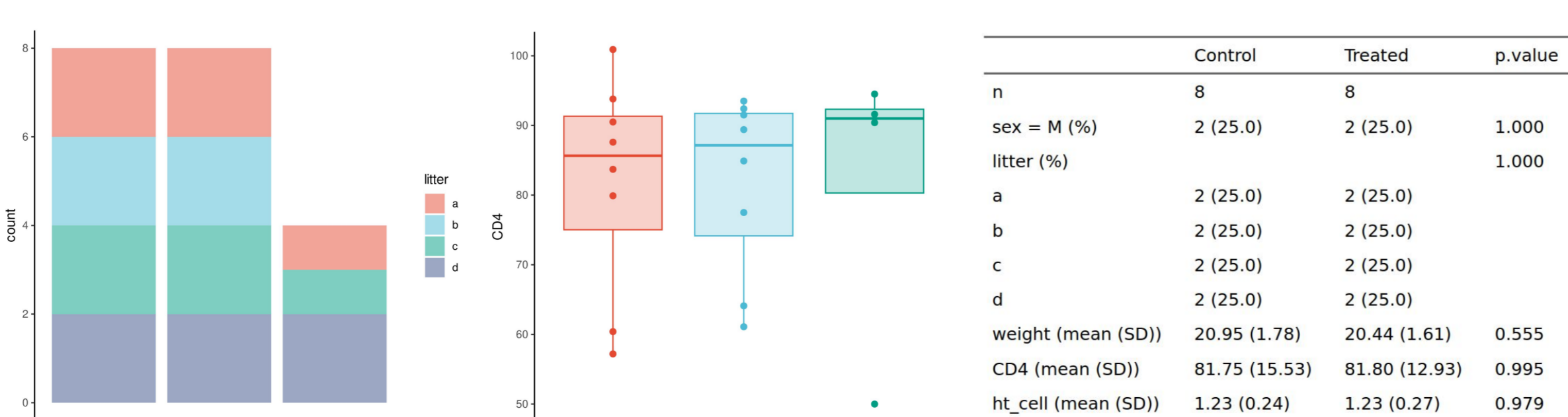### Experiments with animals

Evaluation of a treatment on infected mice

Mice were infected with a pathogen to asses the effects of a given treatment.
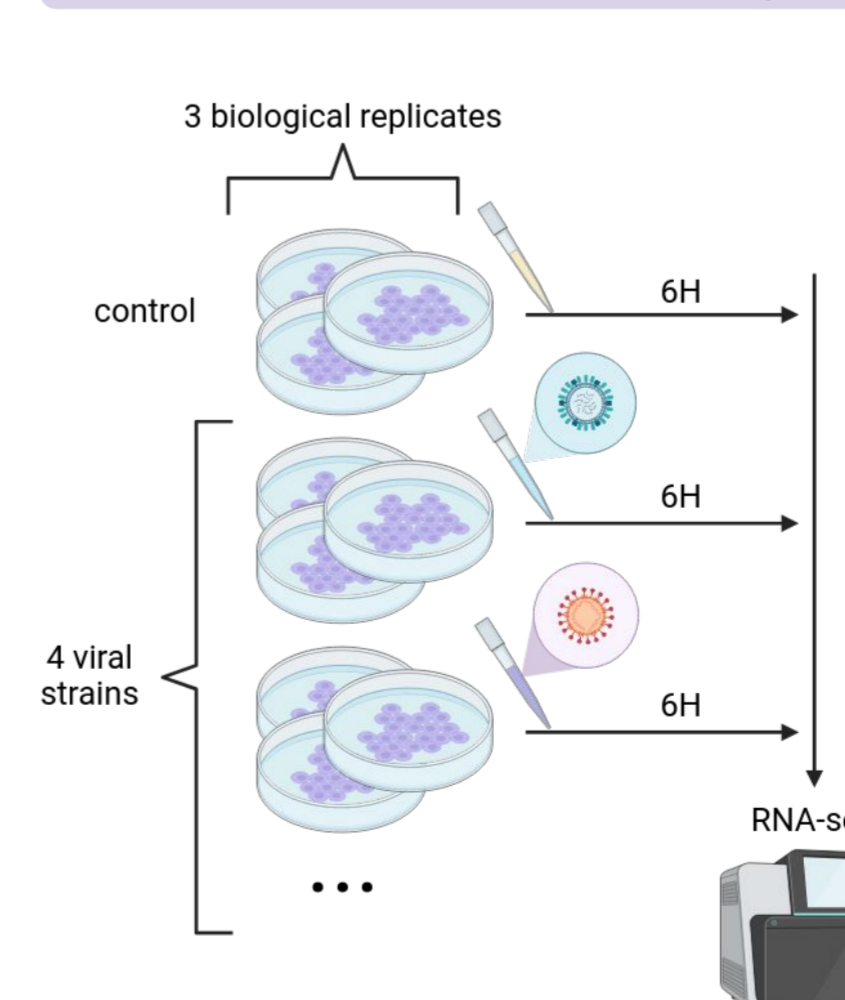
Several litters were reared, how can we distribute the mice into the control and treatment groups ?

Litters reception | Start | | | | | End

Randomisation | Acclimatization | Infection | Injection | Measurement of immune response

Using **Mixnpick**, we generated two groups of 8 mice while taking into account **qualitative** (litters, sex) and **quantitative** (weight, CD4, ht_cells) covariates.

	Control	Treated	p.value
n	8	8	
sex = M (%)	2 (25.0)	2 (25.0)	1.000
litter (%)			1.000
a	2 (25.0)	2 (25.0)	
b	2 (25.0)	2 (25.0)	
c	2 (25.0)	2 (25.0)	
d	2 (25.0)	2 (25.0)	
weight (mean (SD))	20.95 (1.78)	20.44 (1.61)	0.555
CD4 (mean (SD))	81.75 (15.53)	81.80 (12.93)	0.995
ht_cell (mean (SD))	1.23 (0.24)	1.23 (0.27)	0.979

One-way ANOVA test for quantitative variables
Fisher's exact test for qualitatives variables
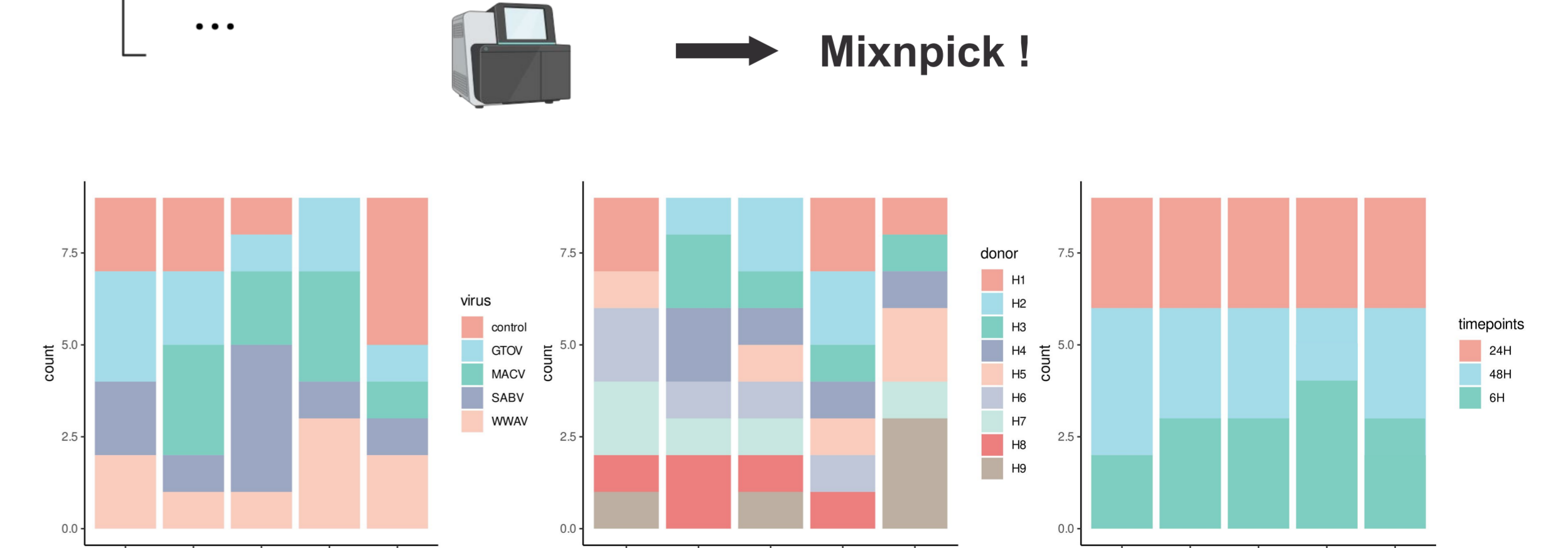
RNA-seq experiment

Comparison of genes expression in hemorrhagic fever patients

Human cells from three donors were collected and infected with different strains of hemorrhagic fever virus. RNA extraction was conducted 6 hours post-infection. The same procedure was repeated with new donors for 24 and 48 hours, resulting in a total of 45 samples.

Considering that one sequencing lane can only handle 10 samples, how can we allocate them while avoiding technical biases ?

→ Mixnpick !

mixnpick.pasteur.fr

INSTITUT PASTEUR