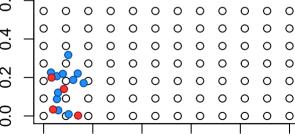simulated dataset