Simulations of Optical Intensities in Microarray Experiments

Daniel J. Arenas

1: Purpose of the Code

This code consists of a new simple theoretical model to simulate optical signals from microarrays searching for multiple organisms. A non-biased search for multiple organisms with multiple probes per organism presents important challenges. Multiple steps in the analysis algorithms, the use of multiple probes searching for the same organism, and the non-normality of experimental data, make nontrivial the calculation of type I (T1E), type II errors (T2E), and the subsequent statistical power (1 – T2E). The purpose of this program is to generate simulated optical data from microchip arrays (such as PathoChip[1–6]) by two methods: an analytical theoretical method and a method that bootstraps experimental data. This type of simulated signal could be utilized in Monte Carlo simulations to test the T1E and statistical power of various data analysis algorithms.

2: Simulation of optical signal by an analytical model

Function: "simulate_analytical_probes' and "gen_from_lognormal"
File containing the functions: "simulateProbes_Functions.R"

2.1: Simulating random noise

First, we discuss how the program simulates the probe intensities in random signal. These simulations are useful for calculating type I errors and for simulating the background signal from probes of absent organisms. For simplicity, and to avoid confusion with other terms that will be introduced shortly, we will refer to the measured signal for each probe as its intensity.

To simulate an optical intensity at each probe, we take into consideration that the intensity must always be positive and therefore cannot be simulated by a normal distribution. Here we will use a chi-squared distribution, which is a sum of squares from sampled normal distributions. This is a computationally-cheap and validated choice for modeling of optical signals.[7] Computationally, the distribution is obtained by first generating random numbers from a Gaussian distribution with a mean of zero and standard deviation of 1:

$$Z = \varphi\,(x) = \; e^{-x^2}.$$ [Eq. 1]

Then each term is squared to obtain the chi-square distribution:

$$P(I) = \; \chi_1^2(I)$$, [Eq. 2]

a distribution with a mean of one. We will use this distribution to model the intensity at each probe expected from random noise.

It should be mentioned that the most general case of the intensity distribution should have the standard deviation as a variable. Here for simplicity, we will parametrize the functions to standard normal distributions so that the unit of intensity corresponds to the mean of [Eq. 2]. And from now on, all intensity units will have the average random (or background) noise as one.

To generate a data-frame for *n* subjects, and *pr* probes each, the function *simulate_analytical_probes* can be used as follows:

> simulate_analytical_probes(pr, n, 0, 0)

The last two parameters in the function refer to optical signal beyond random signal and will be explained in the next sections. Figure 1A and 1B shows representative simulations of random background noise for 100 probes. The figure also denote the probe-average and the 95-percentile cutoffs analytically calculated from the distribution.

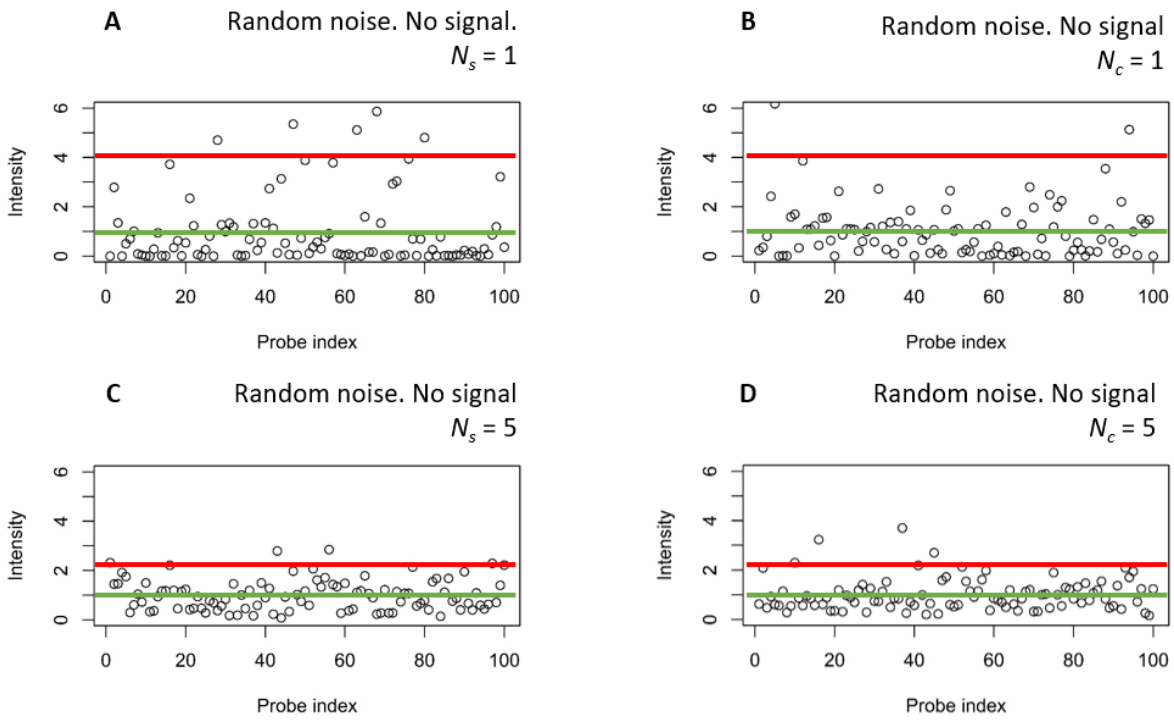

Figure 1. Simulation of random noise. The *x*-axis indexes different probes and the *y*-axis shows the simulated intensity. The top simulations, A-B, correspond to intensities from only one individual in each group (subject, and control) [Eq 2]. Figures C-D show simulations where each probe was averaged across 5 individuals in each group [Eq 2]. For all subfigures, the green line denotes the average intensity and the red line denotes the 95% cutoff expected from the chi-distribution. The simulations demonstrate the intuitive result that as more individuals are averaged for each probe the lower the 95% cutoff becomes. This is expected since fluctuations for each probe should average out across different individuals. A corresponding statement is that the standard-deviation/average of the intensity decreases for the probes whose intensities are due to random noise only. Therefore, the type I error for each probe is fixed regardless of the number of individuals in the group.

The program does not average the signal over the number of individuals as this could be a choice later on in whichever analysis algorithm is chosen to be tested. However, it is useful to mention that averaging the signal over *k* individuals yields a weighted chi-squared distribution with *k* degrees of freedom:

$$P_k(I) = \frac{\chi_k^2(I)}{k}, \text{ [Eq. 3]}$$

that also has a mean of one. Figures 1C and 1D show representative simulations for the averaging of five individuals in a group.

Lastly, it should be emphasized that modelling positive-only signals is not only useful for optical signals from fluorescent probes, but also to other positive-only numbers such as measurement of the number of amplified transcripts.

2.2: Modeling signal from each individual with existing organism

The next consideration is how to simulate signal from an individual that does contain an organism. We will refer to such an individual as a "positive-subject". To simulate the intensity of each probe, we must first consider that:

- One, the intensity fluctuates due to randomness in source, detector, amplifiers (or equivalents).
- Two, although a probe has a non-zero probability of having an intensity below the random noise average, the average of the probe-intensity probability distribution must be larger than that of random noise (Eq. 2) and never less. A sampling distribution may stochastically yield experimental values below, but the probability distribution in the model must have a mean value above the random noise average.
- Three, the increase over the random noise average must be allowed to vary between individuals in the group.

We can model the first two requirements by using a non-central chi-squared distribution. For each individual, i, we first we generate a distribution of non-central normal distributions:

$$Z = \varphi\,(x, y_i) = \; e^{-(x-y_i)^2}. \text{ [Eq. 4]}$$

After the non-central normal distribution is computationally generated, we square each term to obtain a non-central chi-squared distribution:

$$P(I, y_i) = \; \lambda(I, y_i), \text{ [Eq. 5]}$$

where the average intensity is:

$$< I_i >_{probe} = 1 + y_i^2 \text{ , [Eq. 6]}$$

where $<>_{probe}$ denotes averaging over the probes. The units of intensity in the above equation are such that one equals the average intensity generated from random noise.

2.3: Modeling variation across the individuals and the probes searching for the same organism

Variation across individuals can be obtained by varying y_i in Eq. 6; here we will use the log-normal distribution,[8] that is always positive:

$$P(y) = \frac{1}{ys\sqrt{2\pi}} \exp\left(-\frac{(\ln y - m)^2}{2s^2}\right), \text{ [Eq. 7]}$$

and has many applications for modelling particle size dimensions,[9] biological components of shape,[10] virus fitness versus mutation numbers,[11] the number of local lesions in leaves caused by viruses,[12] and the relative abundance of species.[13] The means (μ) and standard deviation (σ) of the distribution function [Eq. 7] are both functions of *m* and *s*. The function *gen_from_lognormal* was written so that the user feeds the desired mean and standard deviations of the log-normal distribution, *P*(*y*) in [Eq. 7]. For our purpose, their meanings are as follows: For each individual *i*, the probe intensity acts as a non-central distribution with an average of $1 + y_i^2$ [See Eq. 6]; across different individuals, the signal above background (y_i^2) varies such that y_i is from a log-normal distribution with mean (μ) and standard deviation (σ). For simplicity, we will incorporate the variance of the probes into the log-normal distribution. This also accomplishes setting variance of average signal above random noise for probes within the same organism.

To summarize the overall simulation methodology, Figure 3 shows a diagram of how each probe is simulated. Figure 3 shows representative simulations for 100 probes for different effect sizes. The top shows a simulation where the subject group has signal above random error (A) while the control does not (B). Subfigures C and D show simulations where the probes were averaged over five individuals. Figure 3C shows the example of simulations for 100 probes, for five subjects, with a desired average signal above random noise of 1 with 0.1 deviation across individuals. The function would be called as follows:

> *simulate_analytical_probes(100, 5, 1, 0.1)*

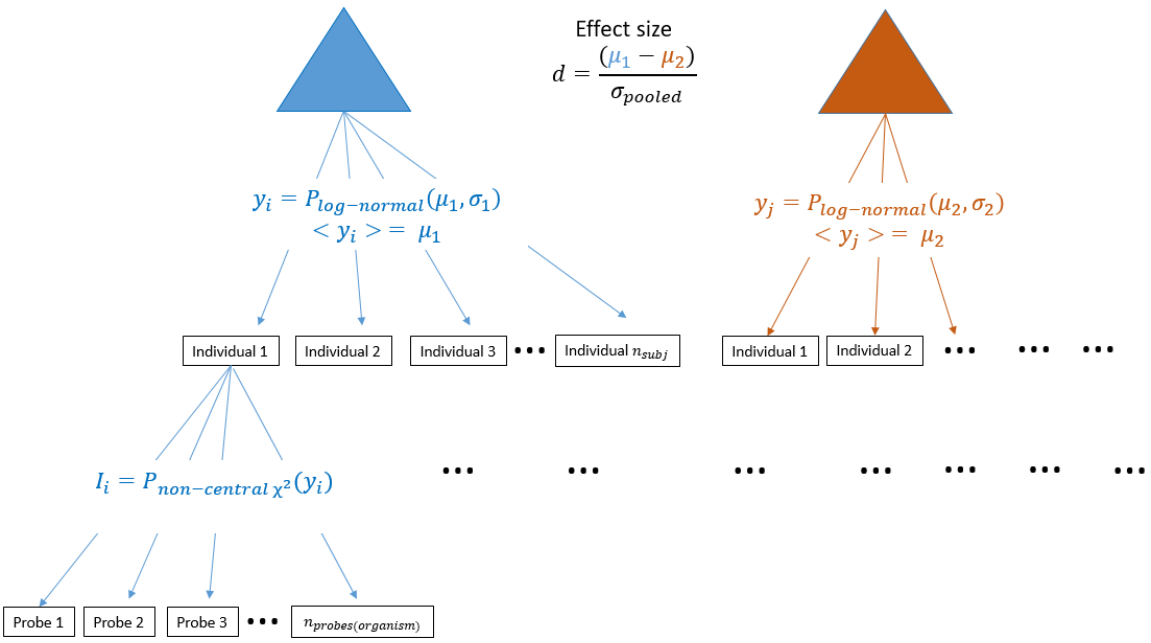

Figure 2. Diagram of general method to simulate data for multiple probes and multiple patients in each group. First an effect size, a difference between the subject and control group, is chosen. Then, the average amount of signal above random noise is simulated for each probe in each patient using the log-normal distribution. Using the aforementioned value, then the intensity of each probe is simulated using the non-central chi squared. The above method ensures that the intensities are always positive, "real" signal is always above random noise, and that there is variance from randomness, variance across probes for the same organism, and variance across individuals.

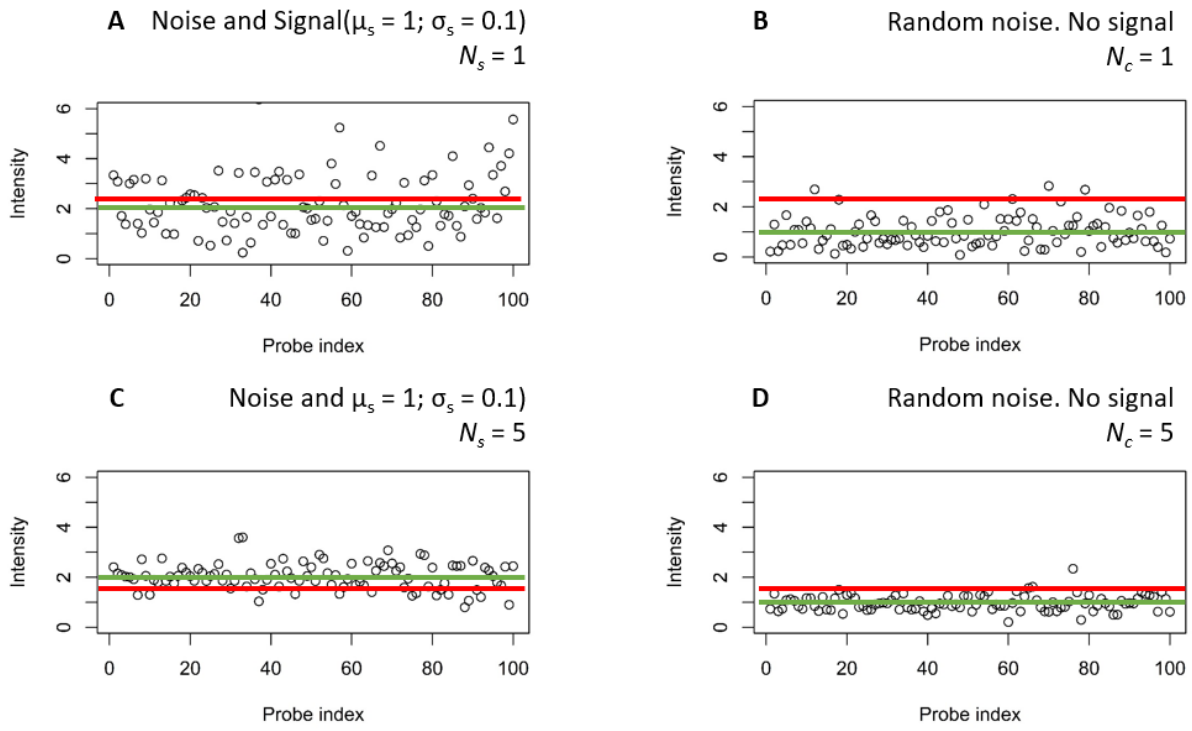

Figure 3. Simulation of signal in a positive subject. (A) Simulated intensities for 100 probes in one patient with signal above random noise. **(B)** Simulated intensities for one patient in the control group (random noise only). Simulation of intensities average over five individuals in the subject group are shown in (C), and for the control group in (D). For all subfigures, the *x*-axis indexes different probes and the *y*-axis the intensity. The green line denotes the probe-averaged intensity and the red line denotes the 95% cutoff expected from the chi-distribution for the number of individuals whose intensities that were averaged. The mathematical models chosen to simulate the intensity allow fluctuation due to randomness in source, detector, amplifiers; each probe has a non-zero probability of having an intensity below the random noise average but their average is higher; the increase over the random noise can vary from individual to individual.

2.4: Discussion of a potential effect size metric

Calculating the type II error (and subsequent statistical power) requires an a-priori effect size – a magnitude of the difference between subjects and controls.[14] We did not find any reports in the literature that simulate effect sizes for simulations of pathogen searching equipment (or any target DNA/RNA). As there are several stochastic processes (variation of probes across each organism, and variation across individuals), and non-normality of signal, it is reasonable to expect that traditional use of effect size for calculation of statistical power have very limited utility.

For simplicity in discussing differences between subjects and control groups, based on the way we model the signal across different individuals we could use the standardized mean difference as the effect size:

$$d = \frac{\mu_{subjects} - \mu_{controls}}{\sigma_{pooled}}$$,[Eq. 9]

where the numerator represents the mean difference between the groups. This corresponds to μ, the desired average of the log-normal distribution in Eq. 6. The denominator the standard pooled deviation, a function of the standard deviations of the subjects and controls. In our model it is the square of these variables that have units of intensity. It is important to note that this equation cannot be used to calculate statistical power directly – although already a simplified model of optical signal, the mathematical complexity is such that the statistical power should be simulated.

3: Simulation of optical signal from bootstrap of experimental data

Function: "bootstrap_n_pr"

File containing the functions: "simulateProbes_Functions.R"

The purpose of this program is to simulate an optical signal, for a variable number of probes (*pr*) and a variable number of subjects (*n*), from positive control data ("*exp_signal*" vector).

For a chosen number of probes (pr), the signal of each probe is obtained by randomly sampling (with replacement) one of the probes from the experimental data. The process is repeated for the number of subjects (*n*) desired. Then each probe is averaged over all the subjects to obtain one average signal. Output is a data frame where each row corresponds to the data for each subject with one column for each probe.

Author information

Daniel J Arenas: Daniel.arenas@pennmedicine.upenn.edu; danielft77@gmail.com

Conflicts of interest: The authors report no conflicts of interest.

References

1. Baldwin DA, Feldman M, Alwine JC, Robertson ES. Metagenomic assay for identification of microbial pathogens in tumor tissues. *MBio*. 2014;5(5):e01714-14.

2. Banerjee S, Wei Z, Tan F, et al. Distinct microbiological signatures associated with triple negative breast cancer. *Sci Rep*. 2015;5:15162.

3. Banerjee S, Tian T, Wei Z, et al. The ovarian cancer oncobiome. *Oncotarget*. 2017;8(22):36225.

4. Banerjee S, Tian T, Wei Z, et al. Distinct microbial signatures associated with different breast cancer types. *Front Microbiol*. 2018;9:951.

5. Banerjee S, Alwine JC, Wei Z, et al. Microbiome signatures in prostate cancer. *Carcinogenesis*. 2019;40(6):749-764.

6. Seckar T, Lin X, Bose D, et al. Detection of Microbial Agents in Oropharyngeal and Nasopharyngeal Samples of SARS-CoV-2 Patients. *Front Microbiol*. 2021;12:454.

7. Humblet PA, Azizoglu M. On the bit error rate of lightwave systems with optical amplifiers. *J Light Technol*. 1991;9(11):1576-1582.

8. Aitchison J, Brown JAC. THE LOGNORMAL DISTRIBUTION, WITH SPECIAL REFERENCE TO ITS USES IN ECONOMICS,. Published online 1969.

9. Epstein B. The mathematical description of certain breakage mechanisms leading to the logarithmico-normal distribution. *J Frankl Inst*. 1947;244(6):471-477.

10. Darroch JN, Mosimann JE. Canonical and principal components of shape. *Biometrika*. 1985;72(2):241-252.

11. Sanjuán R, Moya A, Elena SF. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc Natl Acad Sci*. 2004;101(22):8396-8401.

12. Kleczkowski A. The transformation of local lesion counts for statistical analysis. *Ann Appl Biol*. 1949;36(1):139-152.

13. Williams CB. The use of logarithms in the interpretation of certain entomological problems. *Ann Appl Biol*. 1937;24(2):404-414.

14. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Routledge; 2013.