Monte Carlo simulations of Pathochip probe-intensities and analysis algorithms

Daniel J. Arenas,[1] Johnson Khor,[1] Patricia Tsao,[1] Zhi Wei,[3] Erle Robertson,[1,2] David C. Fajgenbaum[1]

[1]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2]Department of Microbiology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA

1: Simulation of data

1.1: Simulating random noise

First we show how we simulated probe-intensities when only random signal is expected. 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.[1] Computationally, the distribution is obtained by first generating random numbers from a Gaussian distribution with a mean of zero and standard deviation of 1:

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

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

$$P(I) = \; \chi_1^2(I) \text{ [Eq 2]}$$

a distribution with a mean of one. 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. 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 3]. And from now on, all intensity units will have the average random (or background) noise as one. Figure 1 shows representative simulations of random background noise for 100 probes. The figures also denote the probe-average and the 95-percentile cutoffs from the originating distributions.

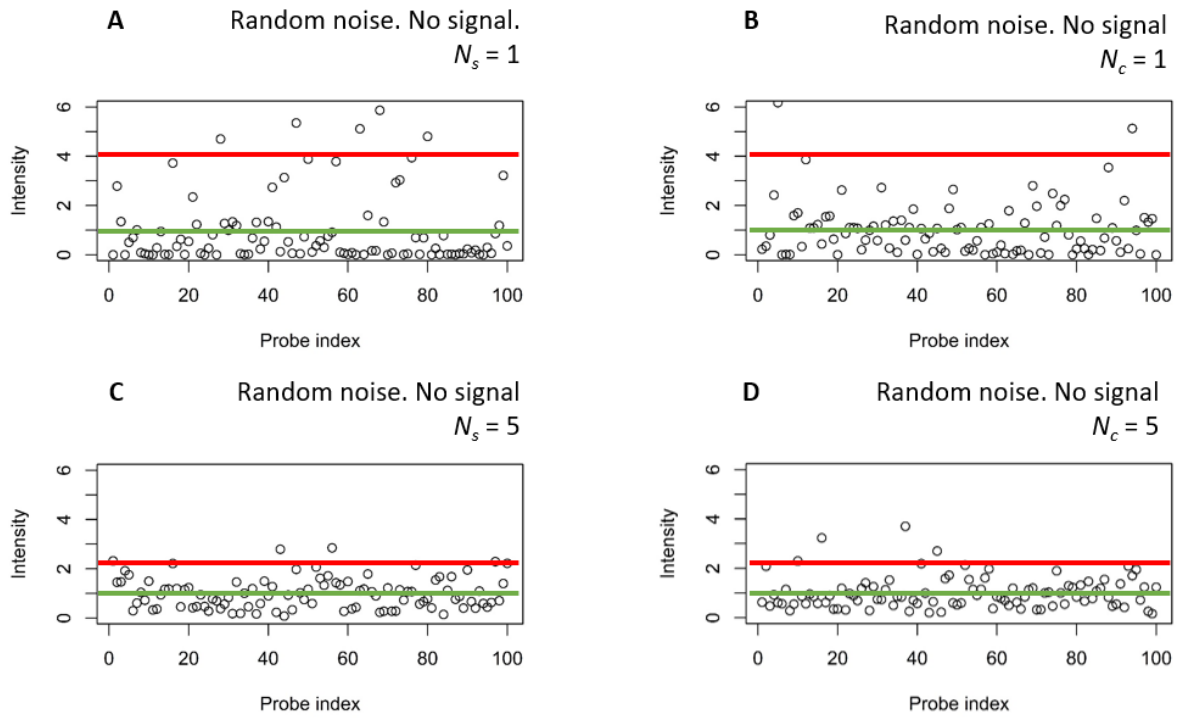

Figure 1. Simulation of random noise. The *x*-axis indexes different probes and the *y*-axis shows the simulated intensity. The top simulations, A-B, correspond to intensities from only one individual in each group (subject, and control) [Eq 2]. Figures C-D show simulations where each probe was averaged across 5 individuals in each group [Eq 3]. 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.

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.

1.2: Simulating an effect size between subjects and controls

Calculating both the type II error and the statistical power requires an a-priori effect size – a magnitude of the difference between subjects and controls.[2] We did not find any reports in the literature that simulate effect sizes for simulations of pathogen searching equipment (or any target DNA/RNA). Here, to model a difference between subjects and controls, we will use the standardized mean difference as the effect size:

$$d = \frac{\mu_{subjects} - \mu_{controls}}{\sigma_{pooled}}$$,[Eq. 4]

where the numerator represents the mean difference between the groups and the denominator the standard pooled deviation, a function of the standard deviations of the subjects and controls. The square of these variables will have units of intensity as it will become clearer later on when we discuss how they are input into the intensity simulations. The amount of target (organism DNA/RNA or other) varies between individuals in the subject group, but we expect the average to be higher than that of the control. In the ideal setting, the amount of target in the control group is zero. For simplicity, we will make this assumption.

1.3: Modeling signal from each individual with existing organism

The next consideration is how to simulate the signal from an individual that does contain the organism. We will refer to such an individual as a "positive-target". To simulate the intensity of each probe, we must first take into account 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 distribution must be larger than that of

random noise (Eq. 2) and never less; 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:

$$Z = \varphi\ (x, y_i) = \ e^{-(x-y_i)^2}. \text{ [Eq. 5]}$$

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

$$P(I, y_i) = \ \lambda(I, y_i), \text{ [Eq. 6]}$$

where the average intensity is:

$$< I_i >_{probe} = 1 + y_i^2 \text{ ,[Eq. 7]}$$

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.

1.4: Modeling variation across individuals and same-organism probes

Variation across individuals can be obtained by varying y_i in Eq 7. Here we will use the log-normal distribution, that is always positive:

$$P(y) = \frac{1}{ys\sqrt{2\pi}} \exp\left(-\frac{(\ln y - m)^2}{2s^2}\right), \text{ [Eq. 8]}$$

and has many biological applications for modelling scales such as length or size[3] or for modeling the virus fitness versus mutation numbers.[4] The means and standard deviation of the distribution function [Eq. 8] are both functions of *m* and *s*, and they are chosen in the simulations to obtain the mean and standard deviation of *P*(*y*) in [Eq. 4]. 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 2 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.

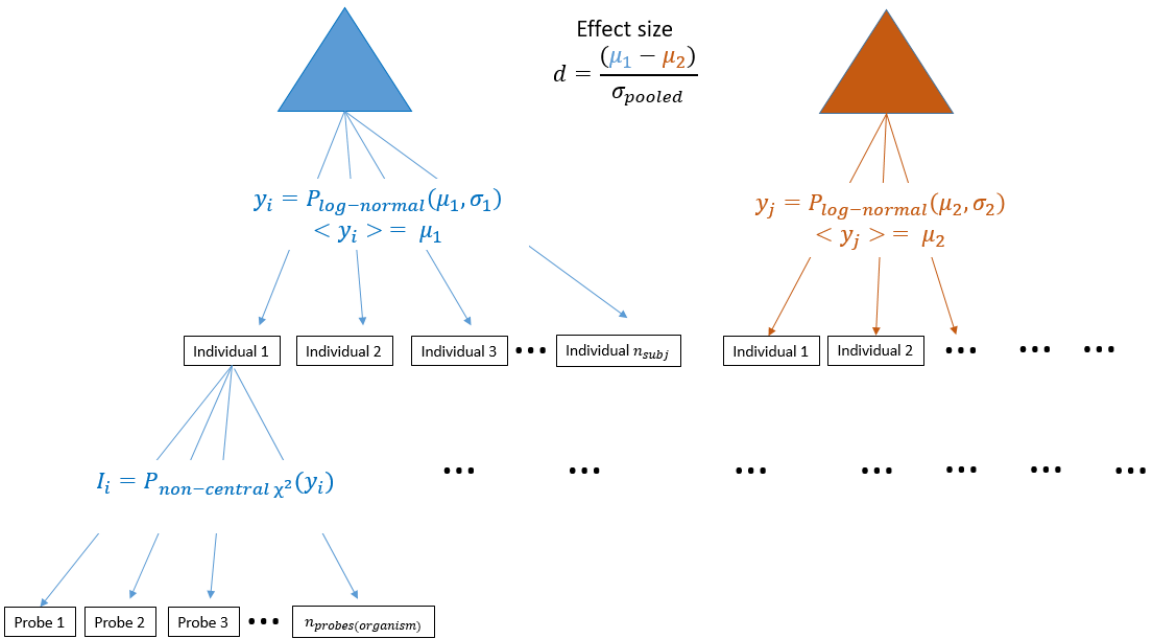

Figure 2. Diagram of general method to simulate data for multiple probes and multiple patients in each group. First an effect size, a difference between the subject and control group, is chosen. Then, the average amount of signal above random noise is simulated for each probe in each patient using the 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.

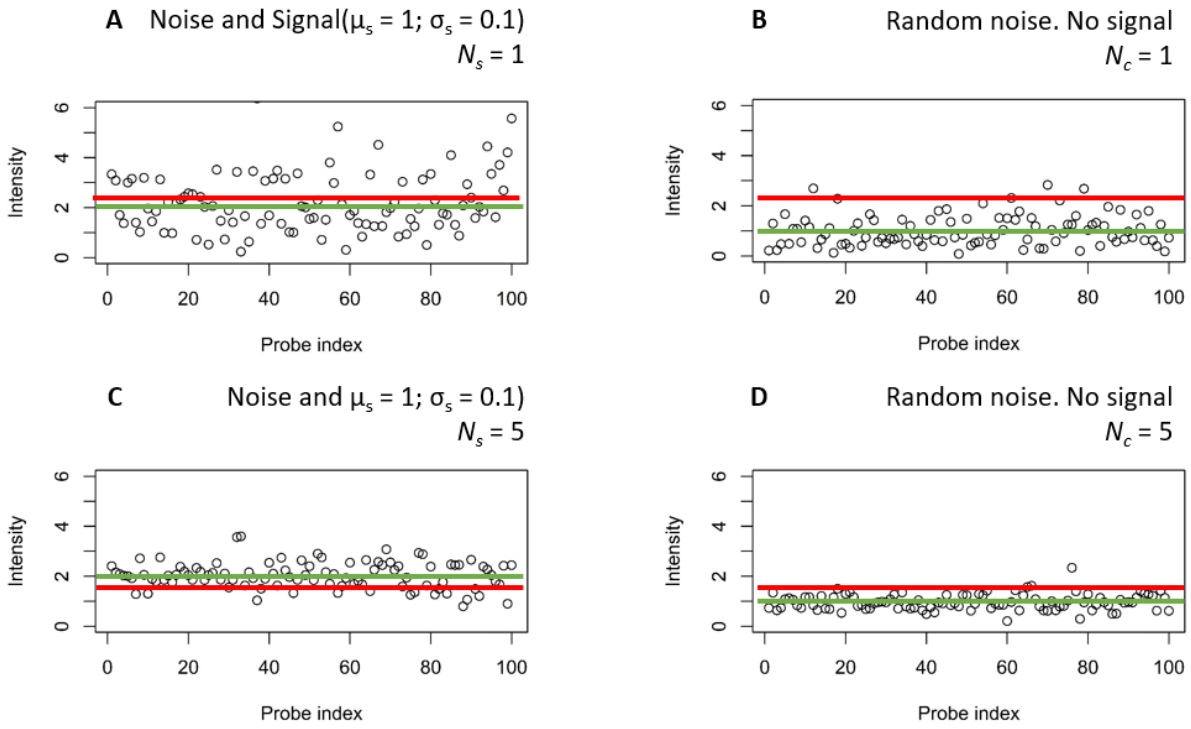

Figure 3. Simulation of signal in the subject group. (A) Simulated intensities for 100 probes in one patient with signal above random noise. **(B)** Simulated intensities for one patient in the control group (random noise only). Simulation of intensities average over five individuals in the subject group 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: Analysis methods

2.1: Choosing a cutoff for positivity

There are different options for choosing a cutoff that decides the positivity of a probe's intensity. Some experiments choose a cutoff based on simultaneous measurement of a control.[5] Here we will choose a cutoff based on the statistics of all probes intensities. This choice is made since it could be straightforwardly applied to any experimental data, by doing a histogram on the experimental intensities

from all probes. This approximation should be very accurate since these experimental techniques simultaneously search for thousands of possible organisms, most of which are not expected to exist in any individual. For our simulations, the cutoff is calculated using the chi-squared distribution with degrees of freedoms equal to the number of individuals in the group (Eq. 3). We will also use this distribution to find the 95% cutoff for when probe intensities from different individuals are averaged.

2.2: The three analysis algorithms

Fig M4-A shows a diagram of the single-probe algorithm, where each probe is treated independently regardless of whether subgroups search simultaneous targets. Each probe is compared to a cutoff value and labeled either above or below. The algorithm can be translated to multiple individuals per group by comparing the average, $\langle I \rangle_{individuals}$, to the cutoff. If the probe is above the cutoff in the subject but not the control, the probe is then considered positive.

Fig M4-B shows a diagram of the probe-average algorithm. This algorithm first pools all probes that search the same target. The probe average $\langle I \rangle_{probe}$ is then compared to the cutoff in both groups. It is worth mentioning that this algorithm is equivalent on whether the probes are averaged first and then averaged over individuals in each group, or the opposite as in $\langle\langle I \rangle_{probes}\rangle_{individuals} = \langle\langle I \rangle_{individuals}\rangle_{probes}$.

Fig M5 shows a diagram of the positive-probe-ratio (PPR) algorithm. First, all probes per organism are pooled. The individual average probe intensity $\langle I \rangle_{individuals}$ is then compared to the cutoff to decide positivity. The ratio of positive probes from subjects and control are then subjected to a binomial-test with a variable significance (unless specified otherwise, 0.05). Please note that the type I error is not trivially synonymous to the significance of the *t*-test as there was a preceding step concerning cutoffs.

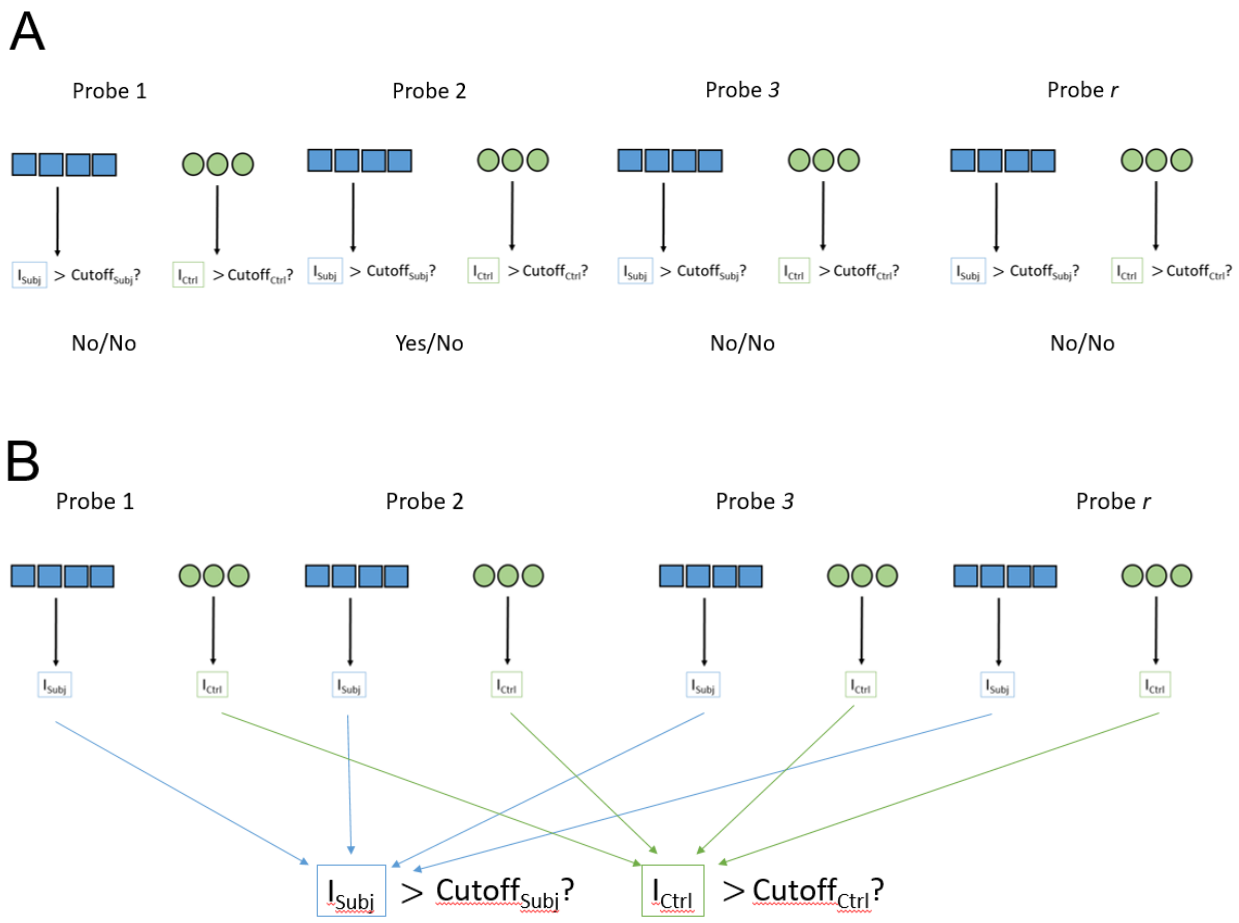

Figure 4. Diagrams of two simple algorithms to analyze multiple-probe data. (A) Each probe can be regarded as independent even if multiple probes search for the same organism. The intensity of each probe is compared to a cutoff value in both the subject and control group ("single-probe" algorithm). **(B)** All the probes for same organism are pooled and their intensities averaged and compared to a cutoff ("probe-average algorithm").

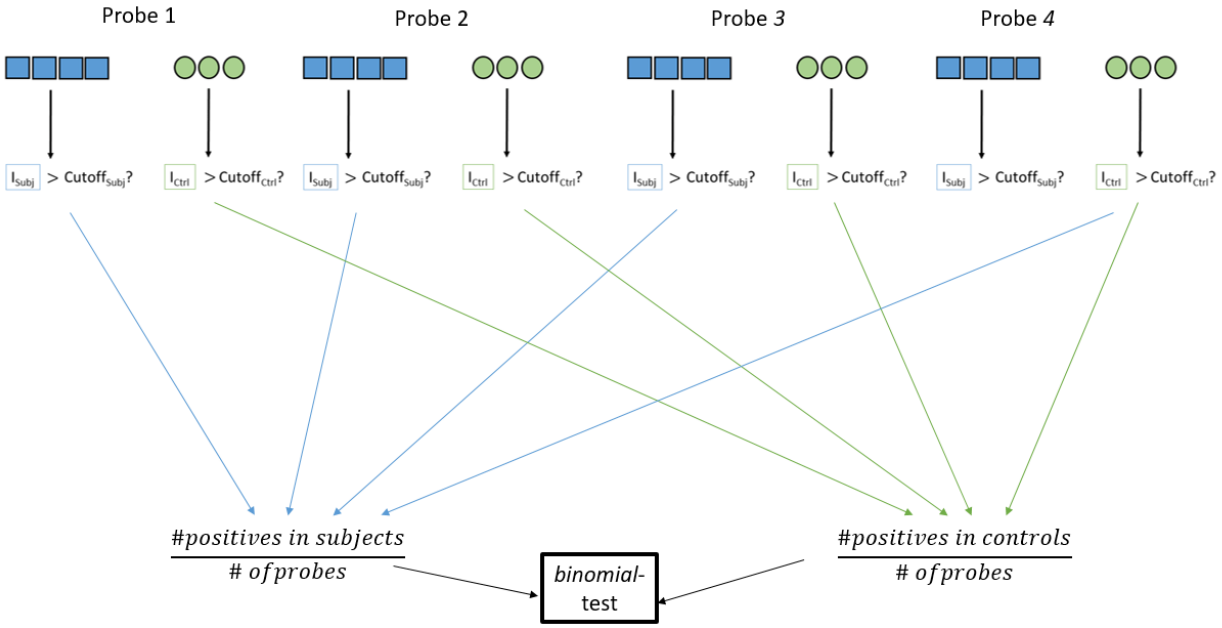

Figure 5. Diagram of PPR algorithm to analyze multiple-probe data. All the probes for one organism are pooled. Each probe is compared to a cutoff. The number of positive probes in each group are recorded. The two proportions are compared to each other by a simple binomial-test. It is important to note the intuitive result that the type I error of this algorithm is not simply the one set for the binomial-test as there is one step before comparing to a cutoff.

Acknowledgements:

DJA would like to thank Lanair Lett for useful discussions and the Gamble Scholarship at Perelman School of Medicine.

Author information:

Daniel J Arenas: Daniel.arenas@pennmedicine.upenn.edu

DJA developed and carried out the ideas to use Monte Carlo simulations to test different analysis algorithms of multiple-probe data, percentile cutoffs for positivity, and adjustment of T1E for reasonable post-validation. DF and DJA developed the positive-probe-ratio algorithm idea. ER's laboratories are the developers of PathoChip. All simulation software was written in *R* by DJA with support from JK. First draft written by DJA and read by all authors.

Conflicts of interest:

The authors report no conflicts of interest.

References:

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

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

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

4. 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.

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

Figure Legends:

Figure M1. 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 3]. 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.

Figure M2. 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 M3. Simulation of signal in the subject group. (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.

Figure M4. Diagrams of two simple algorithms to analyze multiple-probe data. (A) Each probe can be regarded as independent even if multiple probes search for the same organism. The intensity of each probe is compared to a cutoff value in both the subject and control group ("single-probe" algorithm). **(B)** All the probes for same organism are pooled and their intensities averaged and compared to a cutoff ("probe-average algorithm").

Figure M5. Diagram of PPR algorithm to analyze multiple-probe data. All the probes for one organism are pooled. Each probe is compared to a cutoff. The number of positive probes in each group are recorded. The two proportions are compared to each other by a simple binomial-test. It is important to note the intuitive result that the type I error of this algorithm is not simply the one set for the binomial-test as there is one step before comparing to a cutoff.

Figure M6. Histogram of probe-intensities experimentally obtained by PathoChip Arrays. Lymph node tissue from one HHV-8 MCD patient (A) and one PTLD patient (B) were measured along a reactive lymph node without malignant nor autoimmune pathology (C). The histograms denote the frequency of different probe-intensities. The bottom graphs represent a zoom-in to better show the cutoffs used to decide positivity. The blue line denotes the cutoff suggested by the experimental data, the red line the 95% percentile of the histogram.

Figure M7. Probe intensities for the HHv-8 accession and the two HHV-4 accessions across the three lymph node samples. The intensities for all probes searching for HHV-8 are shown for the three samples (Reactive, PTLD, and MCD_{HHV-8} lymph node) are shown in (A). (B-C) show the probe intensities for the two EBV accessions across the three samples. The red line denotes the 95% cutoff of all probes measured for each sample. The results show that for the reactive sample, all three accessions had only a few probes above the 95% cutoff; in contrast, the HHV-4 accessions had many probes in the upper 95^{th} percentile of all probes measured for PTLD. A similar result was found for HHV-8 in the MCD_{HHV-8} sample.

Figure R1. Simulation of statistical power for the single-probe algorithm, all probes independent, across a variable number of signal strengths. (A) Probe intensities were simulated across different signal strengths (*x*-axis), and sample sizes [$N_{subject}$, $N_{control}$], to calculate the statistical power (*y*-axis) from using the single-probe algorithm. The figure shows that unless the signal (in units of random noise average) is significantly higher than one, this analysis algorithm has abysmal statistical power with small sample sizes. Furthermore, depending on the cutoff, the type I error will be about 5% for this algorithm which would be logistically useless for post-validation. **(B)** Simulations for a 99.95% cutoff chosen to set the type I error to a reasonable post-validation number of starting with 50,000 organisms. As expected, the statistical power for small sample sizes becomes further abysmal. The simulations results were corroborated for the experimental PathoChip data where adjusting the cutoff for a post-validation reasonable type I error resulted in HHV-4 being negative for PTLD and HHV-8 for $MCD_{HHV\text{-}8}$.

Figure R2. Simulation of type I error and statistical power across a variable number of probes for the probe-average algorithm. (A) Type-I error calculated from applying the probe-average algorithm to simulated random noise in the subject and control groups. Since the algorithm uses the 95-percentile of the individually-averaged intensities for all probes, the type I error is expected to be the same regardless of sample sizes. Based on the cutoff, we also expect the type I error to be near 0.05 for a small number of probes and to decrease as more probes are averaged for the same organism. **(B)** Statistical power calculated from applying the probe-average algorithm to an effect size where the subject group is expected to have a signal one unit above the random noise. For small sample sizes, this effect size results in low statistical power for a higher number of probes. **(C)** Power calculations for an effect size where the signal is four units above the random noise. **(D)** Power calculations for one individual in each group and varing effect size. For this sample size, signal three units above the random noise is required before a higher number of probes per organism is no longer detrimental to the statiscal power.

Figure R3. Simulation of type I error and statistical power across a variable number of probes for the positive-probe-ratio (PPR) algorithm. (A) Type-I error calculated from applying the PPR algorithm to simulated random noise in the subject and control groups. As expected from the cutoff choice, the type I error was the same across different sample sizes. **(B)** Statistical power for an effect size where the signal is one unit above the random noise. The results for a small number or probes are intuitive. Due to the nature of the *t*-test, a low number of probes expected to have low power regardless of signal strength. **(C)**

Type-I error calculations where the significance of the t-test was adjusted for 50,000 organisms. **(D)** Statistical power after Bonferroni correction.