Statistical methods for spike-in concentrations for quality control in single cell experiments with accompanying poolsplit experiments

Wold Group

December 1, 2016

Single cell analysis offers the opportunity to understand biological process at the basic level of molecular biology; transcriptome processes occuring within a single cell. The challenges of creating and analyzing single cell genomic processes are immense, however, as most analyses using the high-throughput sampling technology (RNA-Seq) in the last decade have been tested and optimized for large collections of cells. Thus techniques that have been adopted for such collections, in particular, reliability of biological samples, do not necessarily apply to single cell samples. Also, the experimental setup of single cell analysis offer an opportunity for additional tests of quality control (hereafter QC) that are not present in bulk analysis.

The outline of this procedure is as follows. First, a probablistic framework, based on Marinov et. al. (2014) is presented for a set of runs (a sample), with a proportion of the runs being single cell runs and a proportion being poolsplit (see Marinov et. al. for definitions or see below). The primary difference between Marinov and this development is that the use of ERCC spikes (which were not available in Marinov et. al.) allows the development of a more expansive statistical approach that allows applications to quality control through statistical testing of agreement with null hypotheses which would assume to hold if certain minimum measures of quality obtain. Additionally, the model presented here also allows the direct estimation of

transcript quantity with error estimates, rather than the estimation of intermediate quantities that are presumed to be correlated with transcript quantity, such as FPKM, TPM, etc.

Three statistical procedures relating to quality control and inference are then discussed. The first and second procedures discusses how the validity of this probabilistic framework can be tested using the theory of likelihood ratio tests, with the first procedure testing individual runs in a sample and the second section testing differences between the poolsplit and single cell runs. The third procedure discusses how the uniqure properties of the poolsplit experiment can be used for quality control over the entire dataset. All of these procedures allow statistical tests which can be interpretted through the usual paradigm of statistical inference.

Probabilistic framework

In this section we derive a probability model for the poolsplit versus single cell experiments. This model allows us to calculate the probability of molecular capture in a single-cell experiment, which can be used both for quality control and inference when there is a small number of transcripts in a cell. This probability model depends not on magnitude of the number of reads but rather on whether any of a type of molecule is captured. In addition, for estimation of larger number of transcripts, a regression model is developed using the spike concentrations, and the best model among a number of common measures of RNA-seq results (FPKM, TPM, counts) is selected for inferences.

We consider the case of a combined poolsplit/single cell experiment (hereafter referred to as a "sample"), with n_s single cell tubes and n_p poolsplit tubes, in which a number of spikes of known quantity are added to each tube. The unit of observation is the spike-in per tube, so if there are S spike-ins and T tubes, there are ST observations to be used in the analysis. There is one presumed parameter common to all tubs, p_{smc}, which is the probability of single molecule capture.

Define the following terms for each observation:

$$\begin{aligned} c &= \text{concentration} \\ t &= \text{number of transcripts} \\ \text{success} &= \text{one or more transcripts captured in cell} \\ \text{failure} &= \text{no transcripts captured in cell} \end{aligned}$$

The quantity p_{smc}, which is unknown, is what we are interested in estimating, for multiple reasons. The estimation of p_{smc} can be done by the method of maximim likelihood, for which there is a well-developed statistical methodology for making inferences, including those of means, standard deviations, and joint distributions of estimates from this procedure. The design of the experiment is such that p_{smc} should be the same for all tubes, whether considered individually or partitioned into poolsplit or single cell (since there are 96 ERCC spikes, relatively good estimates can be obtained by estimating p_{smc} for a single tube). Thus there is a well defined null hypothesis of no difference in the p_{smc} that can be tested through usual statistical techniques.

Explicating the above definitions, the concentration is the expected number of ERCC molecules with are injected into the cell (so this expectation is not necessarily a whole number). The number of transcripts injected into the cell is assumed to follow a Poisson distribution with the mean of this distribution being the concentration, so

$$\text{pr}[t] = \frac{c^t e^{-c}}{c!}$$

It is straight-foward to display the probability of successfully capturing one or more transcripts (referred to as success in the above table) for any fixed number of transcripts:

$$\begin{aligned} \text{pr}[\text{success}|t] &= 1 - p_{\text{smc}}^t \\ \text{pr}[\text{failure}|t] &= p_{\text{smc}}^t, \end{aligned}$$

where failure, as defined above, is not capturing any transcripts. To obtain the overall probability of success as a function of the p_{smc}, the law of total

probability is applied, as follows:

$$\begin{aligned}
\text{pr}[\text{success}|p_{\text{smc}}] &= \sum_{t=0} \text{pr}[\text{success}|p_{\text{smc}}, t]\text{pr}[t] \\
&= \sum_{t=0} \left[1 - p_{\text{smc}}^t\right] \frac{c^t e^{-c}}{c!} \\
\text{pr}[\text{failure}|p_{\text{smc}}] &= \sum_{t=0} \text{pr}[\text{failure}|p_{\text{smc}}, t]\text{pr}[t] \\
&= \sum_{t=0} p_{\text{smc}}^t \frac{c^t e^{-c}}{c!}
\end{aligned}$$

It should also be noted that there is a subtle difference in the likelihood between single cell analysis and poolsplit, as the poolsplit go through an additional step of pooling and then splitting. This is not expected to materially affect the results of the LR tests, however.

Using the probabilities derived above, note that the likelihood function is the product of the probabilities of success or failure in the individual tubes, with a successful tube being represented with the probability of success and an unsuccessful tube being represented with the probability of failure, as follows:

$$\text{L}(p_{\text{smc}}) = \prod_{i \in \text{success}} \text{pr}[\text{success}_i|p_{\text{smc}}] \prod_{i \in \text{failure}} \text{pr}[\text{failure}_i|p_{\text{smc}}],$$

where i is the subscript of the particular cell, and the log-likelihood is

$$\ln \text{L}(p_{\text{smc}}) = \sum_{i \in \text{success}} \ln(\text{pr}[\text{success}_i|p_{\text{smc}}]) \sum_{i \in \text{failure}} \ln(\text{pr}[\text{failure}_i|p_{\text{smc}}]),$$

or

$$\ln \text{L}(p_{\text{smc}}) == \sum_{i \in \text{success}} \sum_{t=0} \left[1 - p_{\text{smc}}^t\right] \frac{c^t e^{-c}}{c!} \sum_{i \in \text{success}} \sum_{t=0} p_{\text{smc}}^t \frac{c^t e^{-c}}{c!} \quad (1)$$

QC1: Likelihood ratio (LR) test for $\mathbf{p}_{smc}$ for individual tubes

Statistical hypotheses can now be tested straight-forwardly using likelihood ratio tests. In particular, if $p_{\text{smc}}^{\text{one tube}}$ is the probability of single molecule capture for only one tube, and $p_{\text{smc}}^{\text{all others}}$ is the probability of single molecule capture for all the other tubes, then the LR test is set up as follows. Let $p_{\text{smc}}^{\text{all others}}$ be the probability of single molecule capture for all the tubes except the one being tested, and $p_{\text{smc}}^{\text{one tube}}$ is the probability of single molecule capture for only the one tube, then

$$-2\left(\ln\frac{\mathrm{L}(p_{\text{smc}})}{\mathrm{L}(p_{\text{smc}}^{\text{all others}})\mathrm{L}(p_{\text{smc}}^{\text{one tube}})}\right),$$

has an asymptotic chi-squared distribution with one degree of freedom under the null hypothesis that the one tube being tested has the same p_{smc} as all the other tubes.

Since there are n_s+n_p tubes, there will be LR n_s+n_p tests with associated p-values, the significances of which can be adjusted by any of the usual Bonferroni or FDR techniques. Note that the power of these LR tests (the ability to reject the null hypothesis of equal p_{smc} between the tested tube and the other tubes) will be relatively low with the number of spikes in the ERCC spike set. A test of systematic differences between the poolsplit and single tubes that will have greater power to detect p_{smc}, is discussed in the next section.

QC2: Likelihood ratio (LR) test for $\mathbf{p}_{smc}$ between single and poolsplit tubes

The LR approach outlined above can be modified to test the validity of the poolsplit versus the single tubes. In this case, the denominator of the LR is not one tube versus all the others, but rather the poolsplit tubes versus the

single tubes. The LR test statistics is as follows:

$$-2\left(\ln\frac{\mathrm{L}(p_{\mathrm{smc}})}{\mathrm{L}(p_{\mathrm{smc}}^{\mathrm{poolsplit}})\mathrm{L}(p_{\mathrm{smc}}^{\mathrm{single}})}\right),$$

which once again has an asymptotic chi-squared distribution with one degree of freedom, where the two log likelihoods in the denominator are restricted to observations in either the poolsplit or single tubes, respectively.

QC3: Comparing means and variability from poolsplit and single tubes

One aspect of the poolsplit design that may be utilized for quality control is that of the analysis of increased variation of total transcripts in the single tubes as composed to the poolsplit tubes. This difference in variation can be analyzed through the distribution of the ratio of spike counts to gene counts, since the same amount of spike concentration are put into each tube, whereas the gene counts are presumably reflective of the total transcripts in a tube. That is, it is supposed that if, say, counts are elevated or decreased in a particular tube, they will be elevated or decreased in an equal manner for both spikes and genes, so that

$$\frac{\text{reads in genes}}{\text{reads in spikes}} = \frac{\text{Transcripts in genes}}{\text{Transcripts in spikes}}$$

All of these quantities are known except the number of transcripts in genes, which can of course be estimated once the number of reads are known. So

$$\text{Transcripts in genes} = \frac{\text{reads in genes}}{(\text{reads in spikes})(\text{Transcripts in spikes})},$$

but the transcripts in the spikes is constant (aside from random error) across all tubes, so use of the ratio gene counts to spike counts gives an estimate of the of the total transcripts in genes.

As a measure of quality control, differences in the ratio of spike reads to overal reads between the poolsplit and single cell tubes can be used. The

logic here is that since the poolsplit tubes average the variation in the genic transcripts that arises from normal cellular stochasticity, it would be expected that the means of the poolsplit and the means of the single tubes should be (statistically) equivalent, whereas, due to stochasticity, the variation should be greater in the single tubes. As an example of this phenomenon, assume that X_i has a mean μ and variance σ^2. Then the single cell tubes are n_s realizations from this distribution. The poolsplit tubes, on the other hand, undergo several steps that changes this distribution. First, n_p realizations X_i are added together and then divided into n_p tubes, so the mean of a realization from a tube is μ and the variance is σ^2/n_p. So the distribution for the poolsplit tube under a correctly functioning experimental design has the same mean and a standard deviation $\sqrt{n_p}$ less then the single cell. Two obvious tests then suggest themselves. The first is a t-test of the ratios between the poolsplit and single tubes, and the second is a test of equality of variation between those two types of tubes. It would be expected that the means would be the same and the variance would be greater in the single tubes. Failure of either of these statistical tests casts doubt on the quality of an experiment.

Samples used for examples

Four samples from poolsplit/single cell experiments are used in the following examples of the statistical test described above. One dataset used Homo Sapien Purkinje cells from a subject classified as having Aspergers (Hs_asp_purkinje, hereafter Sample 1), another used Home sapien Purkinje cells from a non-Asperger subject (Sample 2), one from Mouse layer V pyramidial cells (Sample 3), and one from Mouse Purkinje cells (Sample 4).

EQC1: Likelihood ratio (LR) test for $\mathbf{p}_{smc}$ for individual tubes

Partial results from the LR test proposed in QC1 are given in Table EQC1. As discussed in the LR procedure, the likelihood given in (1) is maximized

Table EQC1: Likelihood ratios tests and estimates of p_{smc} by individual tubes (top 4)

			p-values	
Sample	Tube	LR	unadjusted	adjusted
Sample 1	p13855	10.56	0.001	0.042
Sample 1	p13850	10.32	0.001	0.048
Sample 1	s13824	7.8	0.005	0.193
Sample 1	s13833	7.5	0.006	0.228
Sample 2	p13657	10.98	0.000	0.036
Sample 2	p13649	8.1	0.004	0.177
Sample 2	p13650	6.62	0.010	0.403
Sample 2	p13651	4.34	0.037	1
Sample 3	s15283	5.9	0.015	0.439
Sample 3	s15287	3.14	0.076	1
Sample 3	p15360	2.82	0.093	1
Sample 3	s15276	2.06	0.151	1
Sample 4	p15300	9.64	0.001	0.060
Sample 4	s15266	5.6	0.017	0.574
Sample 4	p15303	4.32	0.037	1
Sample 4	p15288	3.6	0.057	1

over all tubes to form the numerator of the likelihood ratio, and then maximization is done over the single tube being tested and all other tubes save that one to form the denominator. The four highest likelihood ratio values for each sample are given in Table EQC1, along with the unadjusted and adjusted p-values. While there are a number of highly significant individual tubes as indicated by the p-values of the likelihood ratio test, given the large number of comparisons for each sample, there are no adjusted p-value's at the .01 level, though Sample 1 has two at the .05 level and Sample 2 has one at the .05 level. Maximization of the likelihood was done by exhaustion by evaluating the likelihood at values of p from .01 to .99, by increments of .01, and choosing the value of p that gave the maximum value of the likelihood.

Table EQC2: Likelihood ratios tests and estimates of p_{smc} by poolsplit/single dichotomy

Cell type	LR value	p_{smc} poolsplit	single	combined
Sample 1	52.96	0.27	0.14	0.19
Sample 1*	36.80	0.27	0.15	0.20
Sample 2	0.0	0.15	0.15	0.15
Sample 3	2.15	0.09	0.07	0.08
Sample 4	2.28	0.13	0.11	0.12

*possible outliers removed
Likelihood ratio test of less than 3.7 is not significant at .05 level

EQC2: Example of LR test for $\mathbf{p}_{smc}$ between single and poolsplit tubes

Results from the LR test proposed in QC2 are given in Table EQC2. As discussed in the LR procedure, the likelihood given in (1) is maximized over the single tubes and the combined tubes and then over all tubes together. In Table EQC2 all three of the estimated p_{smc} (poolsplit, single and combined) are given, plus the LR value. Maximization was done by exhaustion (running p from .01 to .99 by increments of .01), which is why the LR value of Sample 2 is zero. All except Sample 1 had LR values that did not achieve significance at the .05 level. Some outliers were removed but the LR value was still significant at an extreme p-value.

EQC3: Comparing means and variability from poolsplit and single tubes

As discussed in section QC3, t-tests for equality of the means and a test for the equality of variances was run for all the datasets. P-values for these tests are given in table EQC3. All t-tests of the equality of means between the

Table EQC3: Mean and standard deviation equality test p-values for ratio

	Means			Standard Deviations		
	poolsplit	single	T-test	poolsplit	single	Var test
Sample 1	2.83	1.80	0.06271	1.97	1.12	0.02097
Sample 2	0.79	1.19	0.2051	0.27	1.34	2.488e-09
Sample 3	0.30	0.30	0.9797	0.05	0.48	2.924e-05
Sample 4	0.70	0.59	0.2479	0.03	0.09	1.055e-10

poolsplit and single tests pass at the .05 level (though Sample 1 is just barely above that threshhold), whereas all the variance tests of equality are rejected. Sample 1, however, has greater variation in the poolsplit tubes than in the single tubes, contrary to the assumed direction of the difference. Figures 1 through 4 are box plots of these ratios by poolsplit and single tubes, the plots varying by the subset of samples displayed.

Figure 1: All Samples (_s in legend is single, _p is poolsplit)

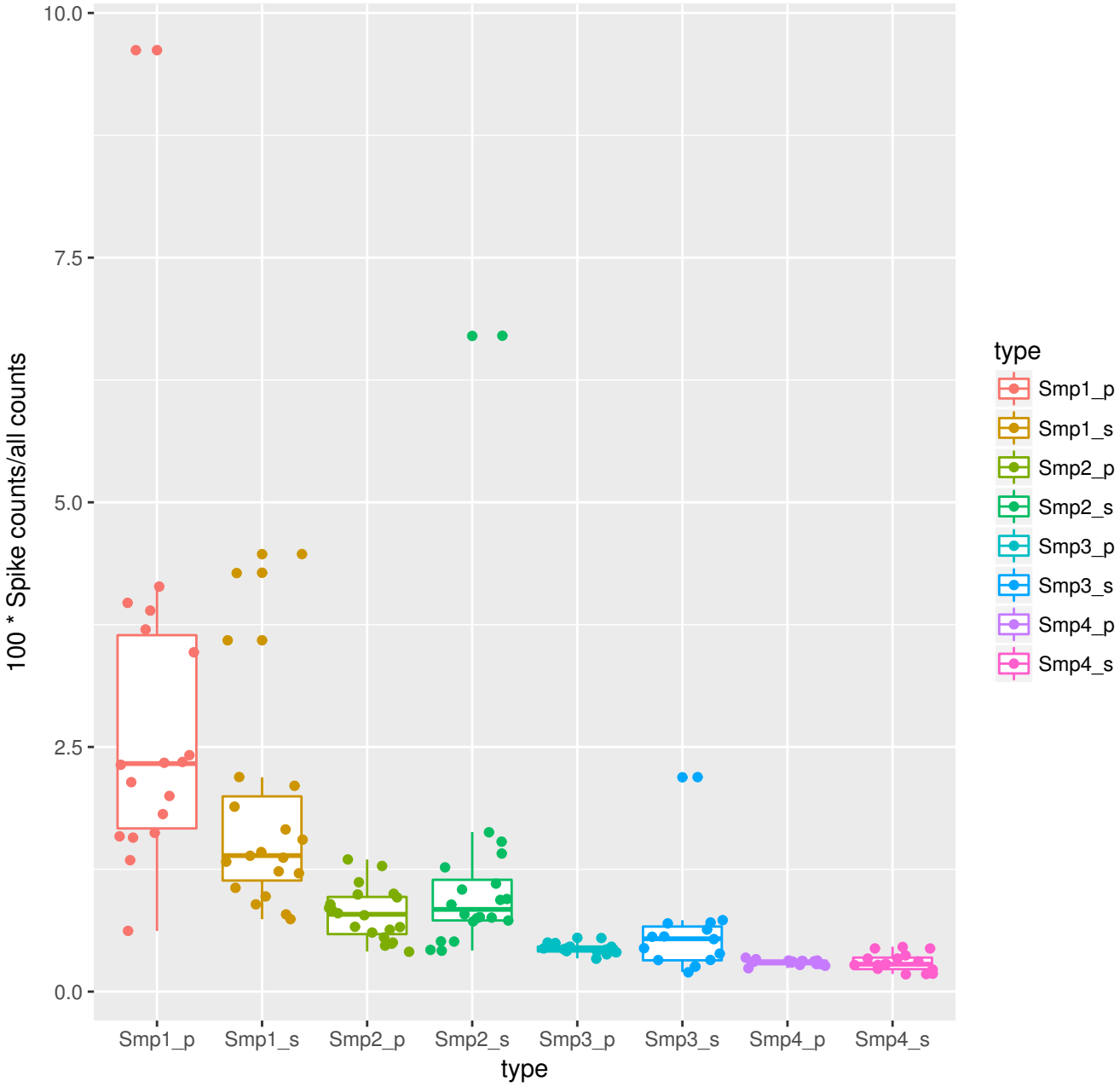


Figure 2: Samples 2, 3, and 4 (_s in legend is single, _p is poolsplit)

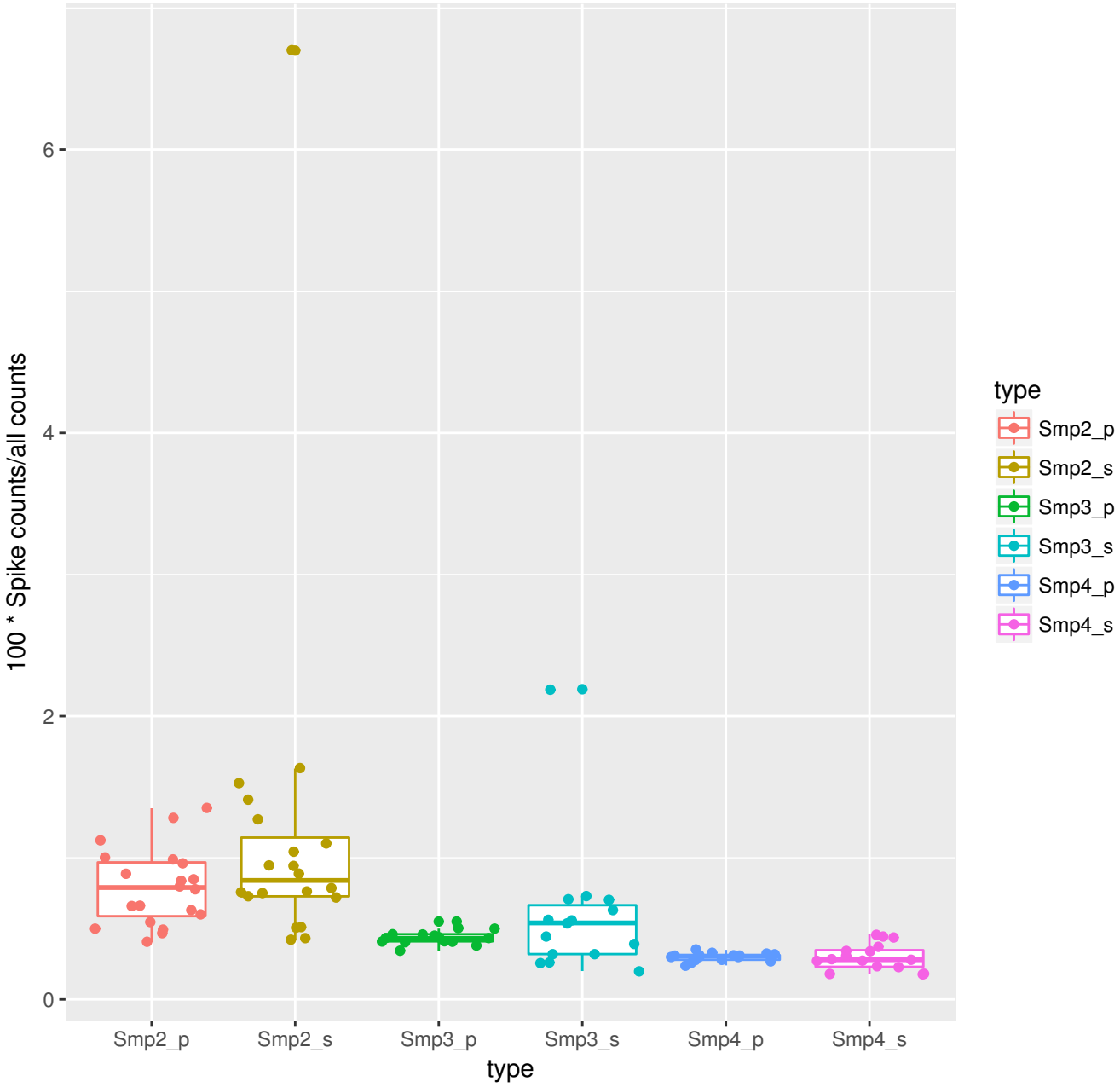


Figure 3: Samples 3 and 4 (_s in legend is single, _p is poolsplit)

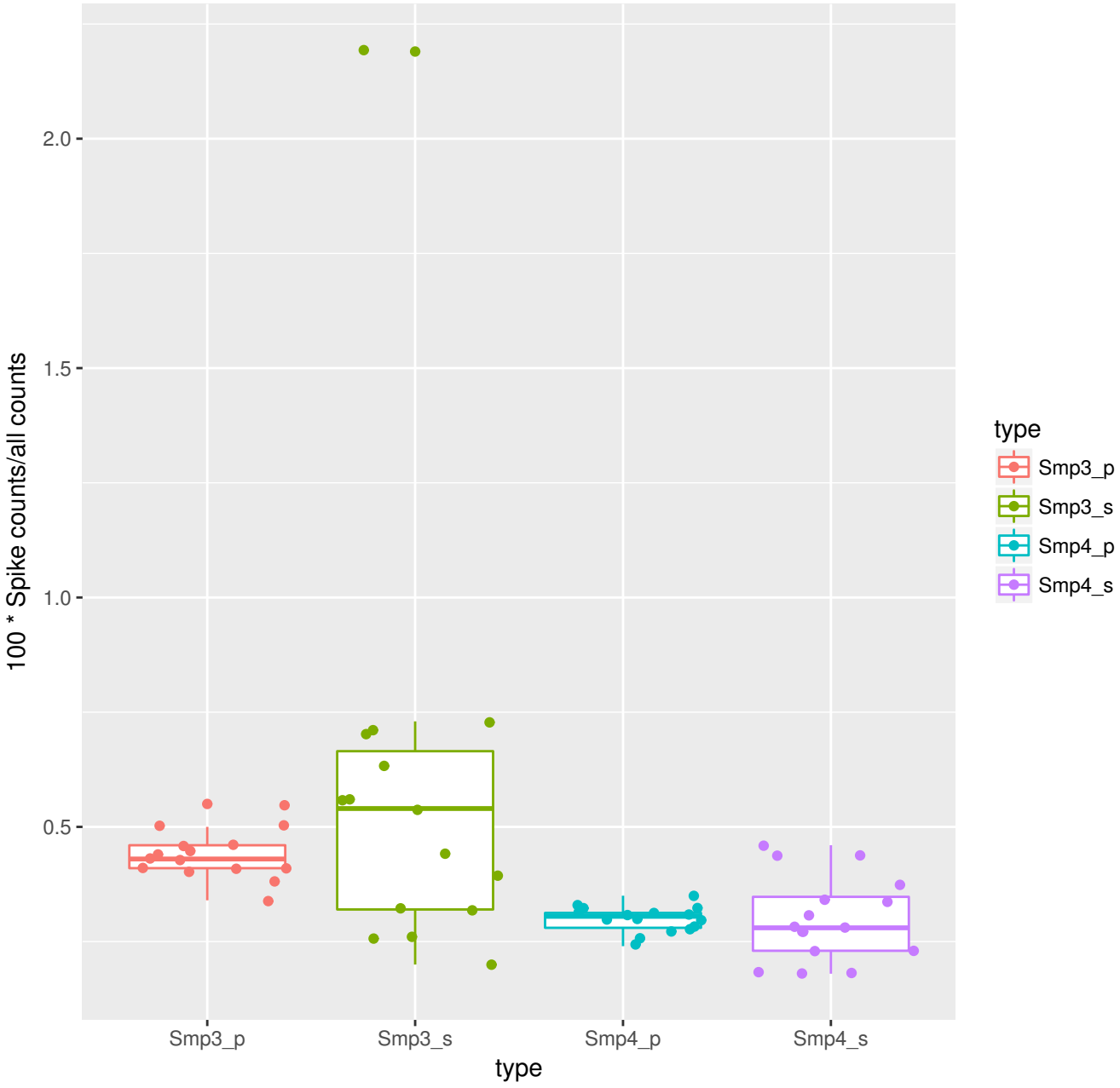
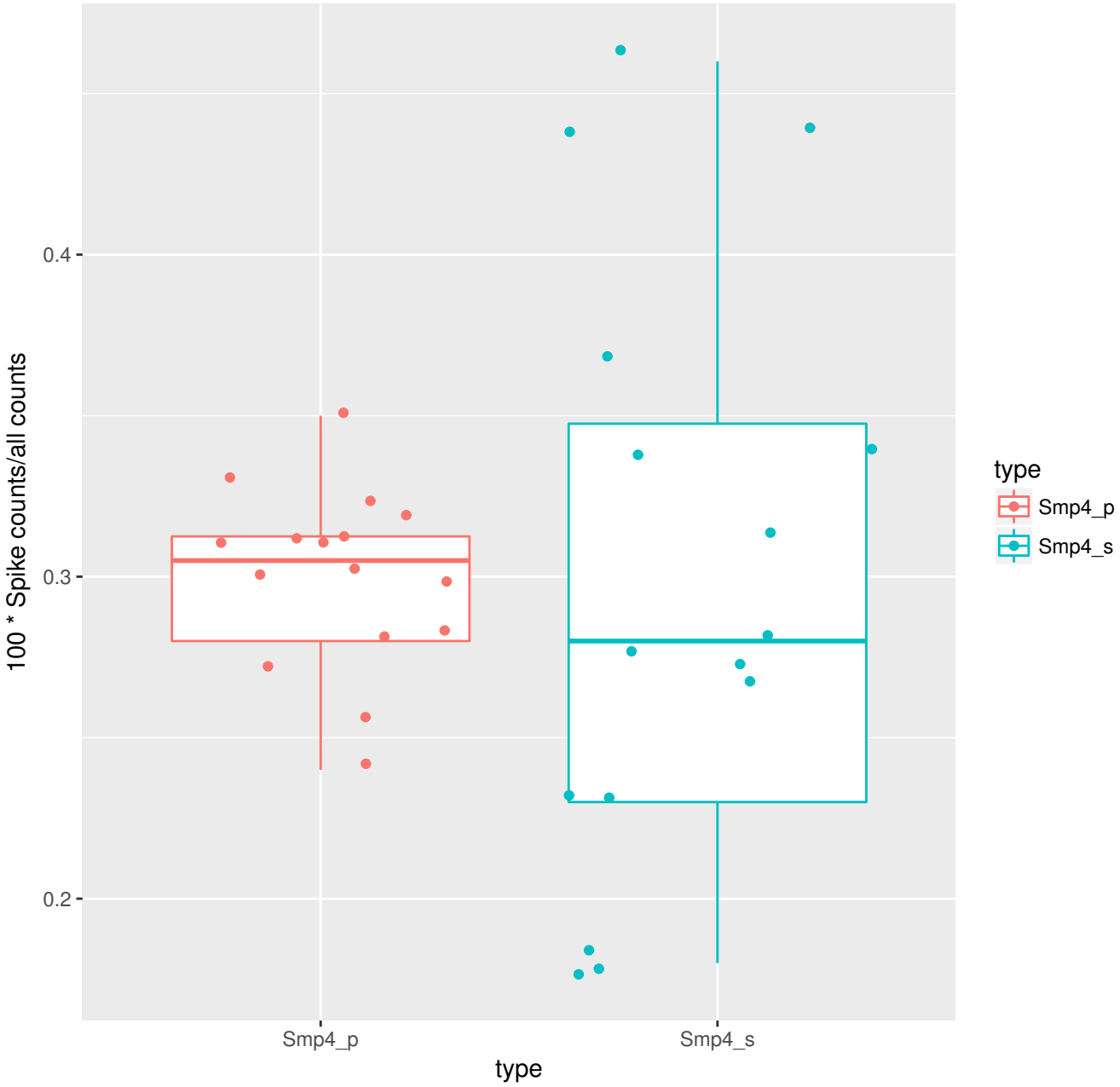


Figure 4: Sample 4 (_s in legend is single, _p is poolsplit)



QC Verdict

Sample 1 fails the equality of p_{smc} between the poolsplit and split runs, using the LR test developed in QC2. The variance of the ratio of spike counts to gene counts (developed in QC3) is also significantly greater for poolsplit rather than single runs. Thus sample 1 should be rejected as a valid sample. Sample 2 has greater variation than one might like but is within acceptable values, as are Sample 3 and 4, which are the highest quality samples.

References

Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ (2014). "From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing" *Genome Research* 3, 496-510.