Pairwise QC pt. 2

Tim

2024-03-11

This writeup focuses on the question of pairwise QC from an empirical perspective. There are two takeaway points.

1. The Poisson-based pairwise QC strategy seems to be reasonable on real data.
2. I find evidence of inflation due to selection bias on the example Gasperini data. The inflation is detectable only for the sparsest pairs.

Application of the Poisson-based QC strategy to the example Gasperini dataset

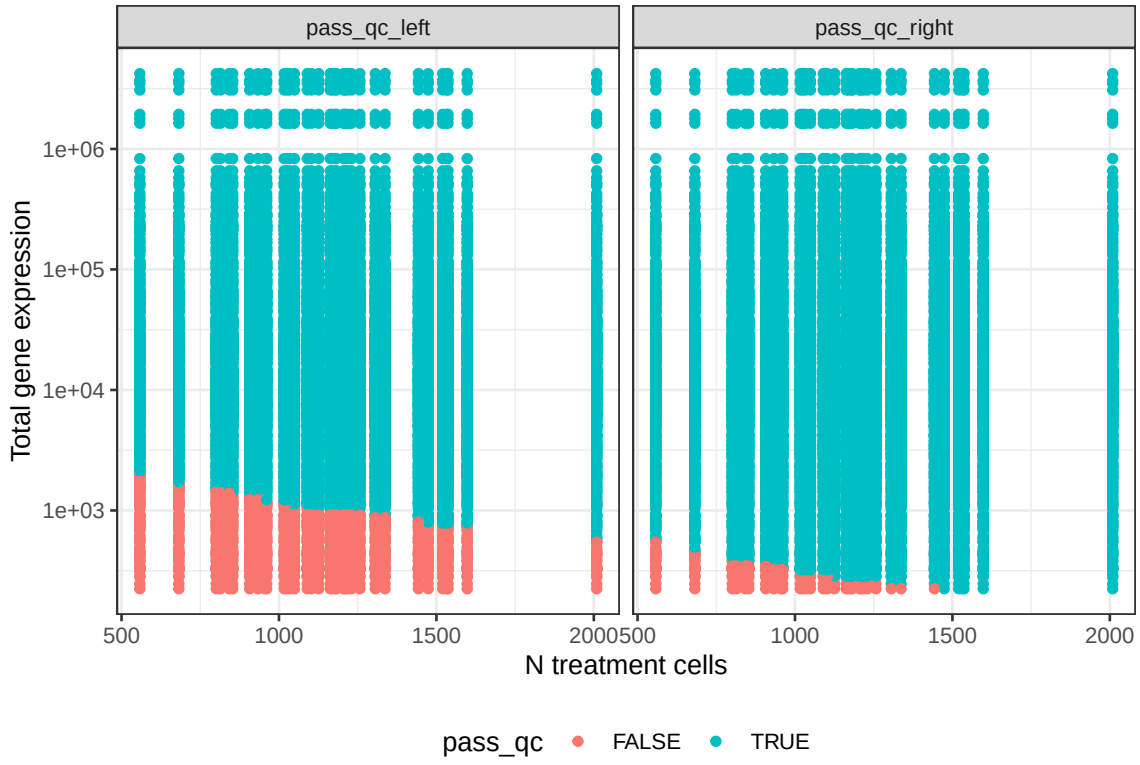
I apply the Poisson-based pairwise QC strategy to the example Gasperini dataset in the `sceptredata` package. To this end, for each *trans* discovery pair, I compute (i) the number of treatment cells (n), (ii) the number of control cells (m), and (iii) the total gene expression across treatment and control cells (t). I set the hypothesized fold change under the alternative (L) to 1/2 for the left-tailed QC and 2 for the right-tailed QC. I compute a "pilot" alternative p-value for each pair using the formula derived in the previous writeup. For example, the left-tailed pilot p-value p_{pilot} for a given pair is

$$p_{\text{pilot}} = \Phi\left(\frac{nt/(m/L+n) - nt/(m+n)}{\sqrt{tmn}/(m+n)}\right).$$

I retain all pairs for which $p_{\text{pilot}} < \alpha$, where I set α to 0.005.

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## [conflicted] Will prefer dplyr::filter over any other package.
```

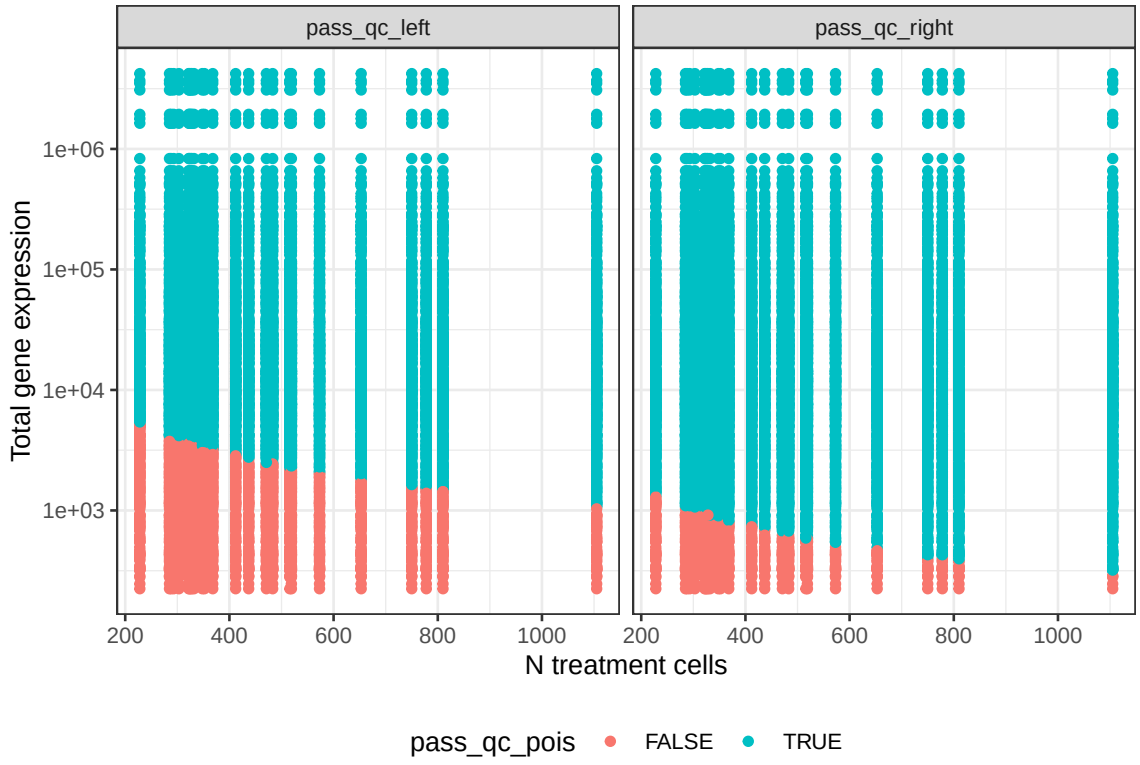
Below, I plot the results of the pairwise QC. Each point represents a pair; the vertical position of a point indicates the total gene expression (t) of that pair, while the horizontal position of a point indicates the number of treatment cells (n) in that pair. I do not plot the number of control cells (m), as the number of control cells is highly similar across pairs and thus does not contain much information. The left-tailed (resp., right-tailed) QC is shown on the left (resp., right).



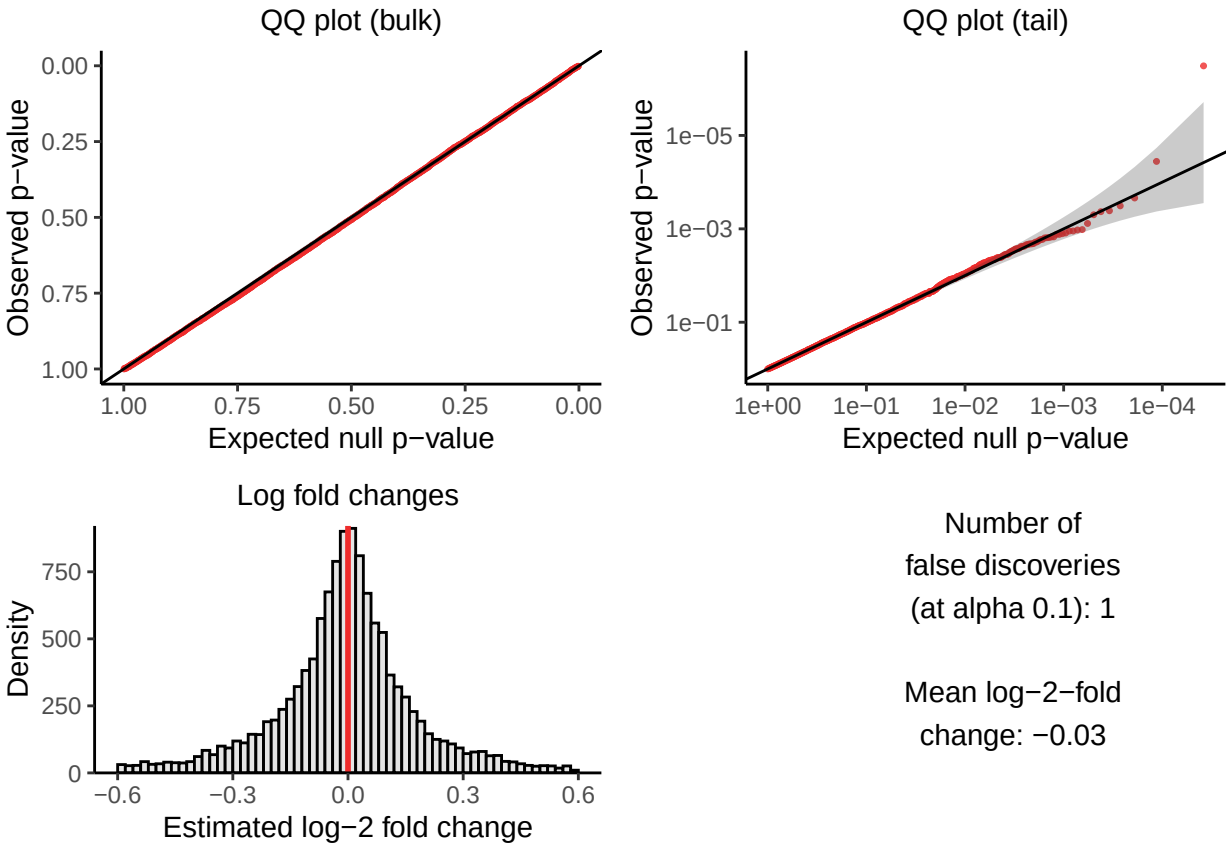
The plot looks reasonable. In particular, a pair is more likely to be retained as its gene expression and/or number of treatment cells increases. Furthermore there are no "outliers;" that is, there is no errant red point (or set of red points) that falls among the blue points (or vice versa).

I repeat this analysis on ~12,000 singleton negative control pairs. I create the same plot for these pairs (below). Again, the plot looks reasonable.

```
## Constructing negative control pairs.
## Note: Unable to generate the number of negative control pairs (20000) requested. Generating as many r
## Generating permutation resamples.
## Running calibration_check in parallel. Change directories to /var/folders/7v/5sqjgh8j28lgf8qx3gbtq1h0
##
```



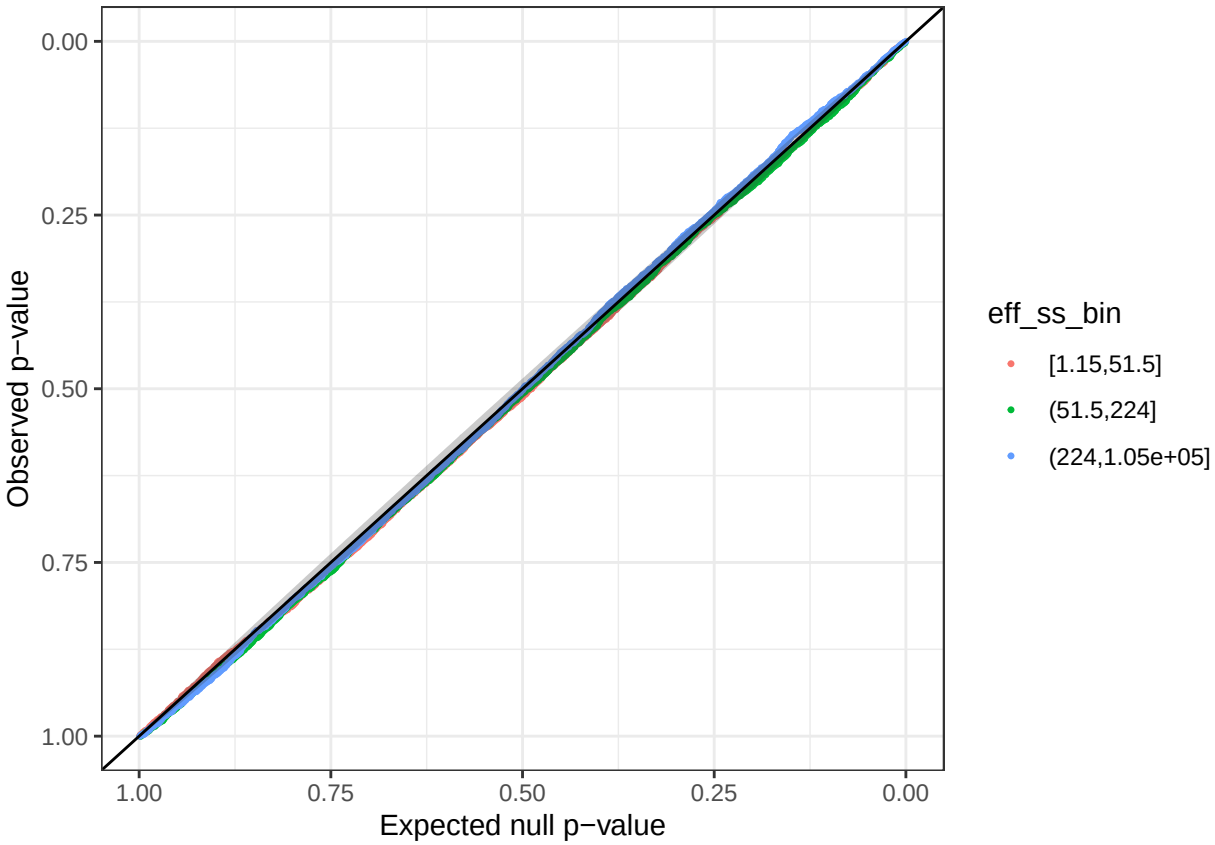
I apply `sceptre` to compute a **right-tailed** p-value for each negative control pair. These right-tailed p-values are uniformly distributed.



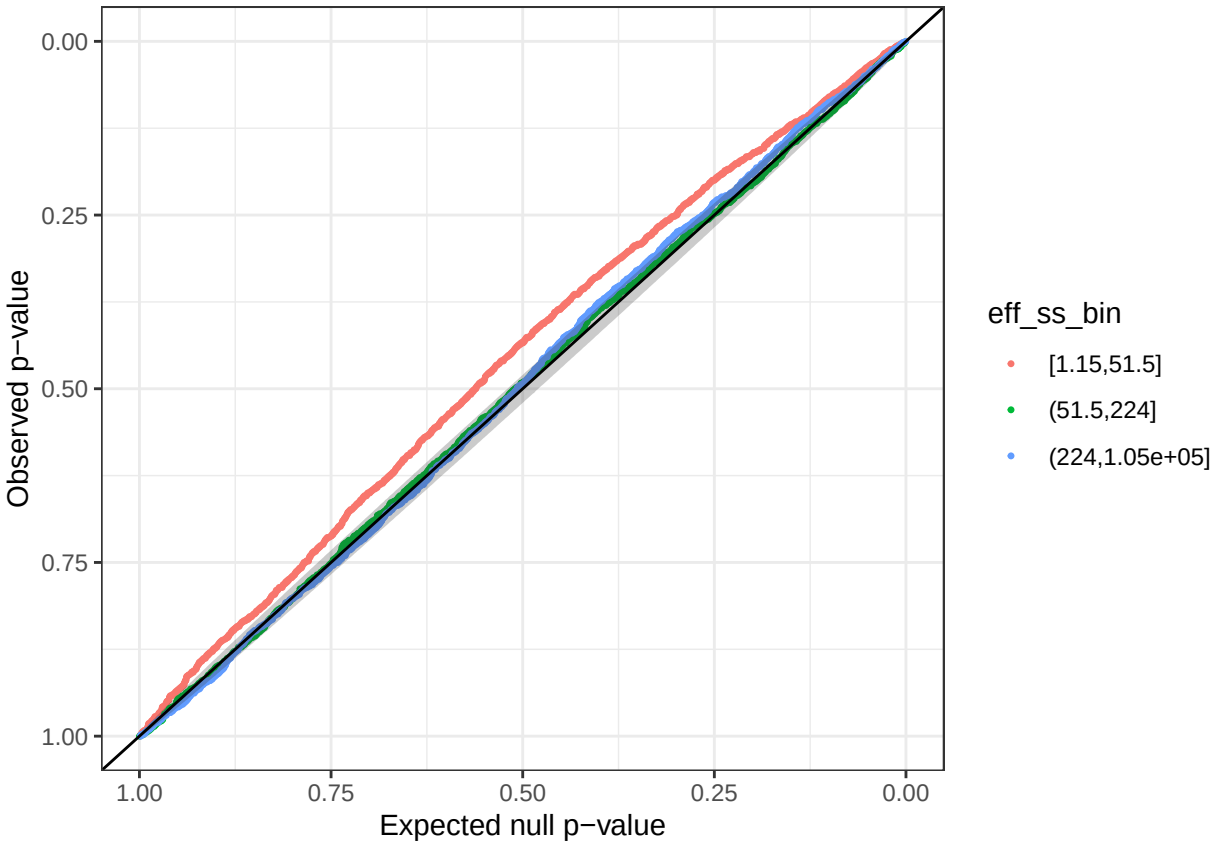
Next, I seek to explore whether conditioning on `n_nonzero_trt` (and `n_nonzero_cntrl`) might induce a selection bias. To this end, following Gene, I partition the pairs into three categories based on their "effective sample size," where the effective sample μ_0 of a given pair is defined as the mean UMI count of the gene in the treatment cells under the null hypotheses, as follows.

$$\mu_0 = \frac{tn}{m+n}.$$

I plot the p-values on a QQ plot, colored by category (red = smallest effective sample size; green = intermediate effective sample size; blue = largest effective sample size). Calibration looks good across all three categories.



Next, I filter on `n_nonzero_trt` and `n_nonzero_cntrl`, retaining pairs for which these quantities exceed 15.



The pairs with a small effective sample size show mild inflation in the bulk of the distribution. This is consistent with Gene's prediction. (Nicely done Gene!)