P and q values in RNA Seq

The q-value is an adjusted p-value, taking in to account the false discovery rate (FDR). Applying a FDR becomes necessary when we're measuring thousands of variables (e.g. gene expression levels) from a small sample set (e.g. a couple of individuals). A p-value of 0.05 implies that we are willing to accept that 5% of all tests will be false positives. An FDR-adjusted p-value (aka a q-value) of 0.05 implies that we are willing to accept that 5% of the tests found to be statistically significant (e.g. by p-value) will be false positives. Such an adjustment is necessary when we're making multiple tests on the same sample. See, for example, http://www.totallab.com/products/samespots/support/faq/pq-values.aspx.

-HomeBrew-

What are p-values?

The object of differential 2D expression analysis is to find those spots which show expression difference between groups, thereby signifying that they may be involved in some biological process of interest to the researcher. Due to chance, there will always be some difference in expression between groups. However, it is the size of this difference in comparison to the variance (i.e. the range over which expression values fall) that will tell us if this expression difference is significant or not. Thus, if the difference is large but the variance is also large, then the difference may not be significant. On the other hand, a small difference coupled with a very small variance could be significant. We use the one way Anova test (equivalent t-test for two groups) to formalise this calculation. The tests return a p-value that takes into account the mean difference and the variance and also the sample size. The p-value is a measure of how likely you are to get this spot data if no real difference existed. Therefore, a small p-value indicates that there is a small chance of getting this data if no real difference existed and therefore you decide that the difference in group expression data is significant. By small we usually mean 0.05.

What are q-values, and why are they important?

False positives

A positive is a significant result, i.e. the p-value is less than your cut off value, normally 0.05. A false positive is when you get a significant difference when, in reality, none exists. As I mentioned above, the p-value is the chance that this data could occur given no difference actually exists. So, choosing a cut off of 0.05 means there is a 5% chance that we make the wrong decision.

The multiple testing problem

When we set a p-value threshold of, for example, 0.05, we are saying that there is a 5% chance that the result is a false positive. In other words, although we have found a statistically significant result, in reality, there is no difference in the group means. While 5% is acceptable for one test, if we do lots of tests on the data, then this 5% can result in a large number of false positives. For example, if there are 200 spots on a

gel and we apply an ANOVA or t-test to each, then we would expect to get 10 false positives by chance alone. This is known as the multiple testing problem.

Multiple testing and the False Discovery Rate

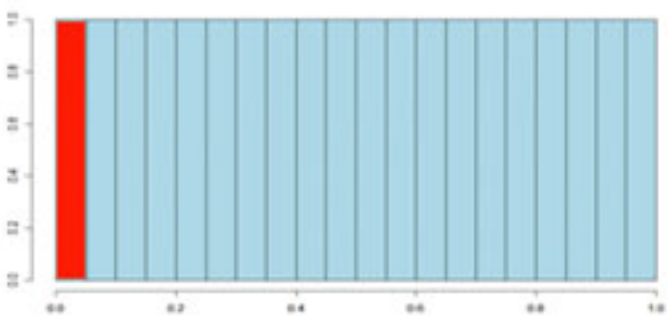
While there are a number of approaches to overcoming the problems due to multiple testing, they all attempt to assign an adjusted p-value to each test, or similarly, reduce the p-value threshold. Many traditional techniques such as the Bonferroni correction are too conservative in the sense that while they reduce the number of false positives, they also reduce the number of true discoveries. The False Discovery Rate approach is a more recent development. This approach also determines adjusted p-values for each test. However, it controls the number of false discoveries in those tests that result in a discovery (i.e. a significant result). Because of this, it is less conservative that the Bonferroni approach and has greater ability (i.e. power) to find truly significant results.

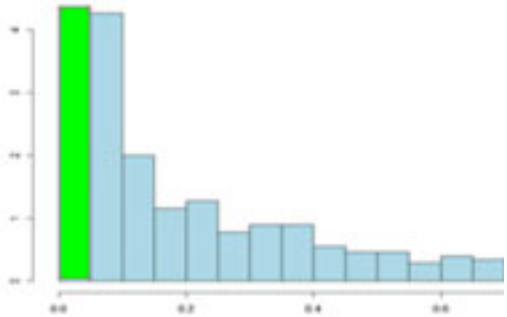
Another way to look at the difference is that a p-value of 0.05 implies that 5% of all tests will result in false positives. An FDR adjusted p-value (or q-value) of 0.05 implies that 5% of significant tests will result in false positives. The latter is clearly a far smaller quantity.

q-values

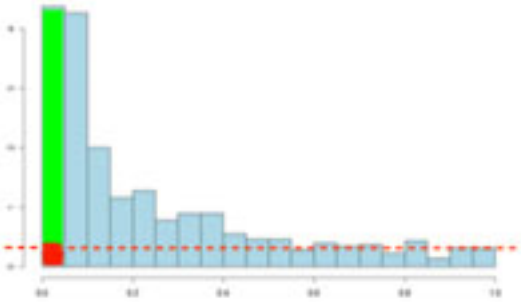
q-values are the name given to the adjusted p-values found using an optimised FDR approach. The FDR approach is optimised by using characteristics of the p-value distribution to produce a list of q-values. In what follows I will tie up some ideas and hopefully this will help clarify some of the ideas about p and q values.

It is usual to test many hundreds or thousands of spot variables in a proteomics experiment. Each of these tests will produce a p-value. The p-values take on a value between 0 and 1 and we can create a histogram to get an idea of how the p-values are distributed between 0 and 1. Some typical p-value distributions are shown below. On the x-axis we have histogram bars representing p-values. Each has a width of 0.05 and so in the first bar (red or green) we have those p-values that are between 0 and 0.05. Similarly, the last bar represents those p-values between 0.95 and 1.0, and so on. The height of each bar gives an indication of how many values are in the bar. This is called a density distribution because the area of all the bars always adds up to 1. Although the two distributions appear quite different, you will notice that they flatten off towards the right of the histogram. The red (or green) bar represents the significant values, if you set a p-value threshold of 0.05.





If there are no significant changes in the experiment, you will expect to see a distribution more like that on the left above while an experiment with significant changes will look more like that on the right. So, even if there are no significant changes in the experiment, we still expect, by chance, to get p-values < 0.05. These are false positives, and shown in red. Even in an experiment with significant changes (in green), we are still unsure if a p-value < 0.05 represents a true discovery or a false positive. Now, the q-value approach tries to find the height where the p-value distribution flattens out and incorporates this height value into the calculation of FDR adjusted p-values. We can see this in the histogram below. This approach helps to establish just how many of the significant values are actually false positives (the red portion of the green bar).



Now, the q-values are simply a set of values that will lie between 0 and 1. Also, if you order the p-values used to calculate the q-values, then the q-values will also be ordered. This can be seen in the following screen shot from SameSpots. Notice that q-values can be repeated.

Ask another question

Rank	Anova (p)	q Value	Power	Cluster
30	0.00436	0.0119	0.993	
77	0.00536	0.0119	0.987	
97	0.00631	0.0119	0.98	
29	0.00655	0.0119	0.979	
43	0.00605	0.0119	0.982	
23	0.0067	0.0119	0.977	
36	0.00632	0.0119	0.98	
28	0.00698	0.0119	0.975	
76	0.00685	0.0119	0.976	
60	0.0067	0.0119	0.977	
10	0.00479	0.0119	0.991	
13	0.00467	0.0119	0.991	
51	0.00432	0.0119	0.993	
91	0.0062	0.0119	0.981	
21	0.00611	0.0119	0.982	
46	0.00414	0.0119	0.994	
45	0.00739	0.0127	0.972	
25	0.00822	0.0137	0.964	
53	0.00903	0.0137	0.956	
6	0.00919	0.0138	0.955	
52	0.01	0.0141	0.946	
2	0.00976	0.0141	0.949	
87	0.0101	0.0141	0.946	
19	0.0109	0.0141	0.938	
96	0.0102	0.0141	0.944	
55	0.011	0.0141	0.937	
50	0.00949	0.0141	0.952	
49	0.0115	0.0144	0.931	
32	0.0127	0.0144	0.918	

To interpret the q-values, you need to look at the ordered list of q-values. There are 839 spots in this experiment. If we take spot 52 as an example, we see that it has a p-value of 0.01 and a q-value of 0.0141. Recall that a p-value of 0.01 implies a 1% chance of false positives, and so with 839 spots, we expect

between 8 or 9 false positives, on average, i.e. 839*0.01 = 8.39. In this experiment, there are 52 spots with a value of 0.01 or less, and so 8 or 9 of these will be false positives. On the other hand, the q-value is a little greater at 0.0141, which means we should expect 1.41% of all the spots with q-value less than this to be false positives. This is a much better situation. We know that 52 spots have a q-value less than 0.0141 and so we should expect 52*0.0141 = 0.7332 false positives, i.e. less than one false positive. Just to reiterate, false positives according to p-values take all 839 values into account when determining how many false positives we should expect to see while q-values take into account only those tests with q-values less the threshold we choose. Of course, it is not always the case that q-values will result in less false positives, but what we can say is that they give a far more accurate indication of the level of false positives for a given cut-off value.

When doing lots of tests, as in a proteomics experiment, it is more intuitive to interpret p and q values by looking at the entire list of values in this way rather that looking at each one independently. In this way, a threshold of 0.05 has meaning across the entire experiment. When deciding on a cut-off or threshold value, you should do this from the point of view of how many false positives will this result in, rather than just randomly picking a p- or q-value of 0.05 and saying that everything with a value less than this is significant.

Conversation of things to consider when setting p values

Hello,

I'm dealing with a classical dilemma: I performed RNA-seq experiment on two biological replicates for condition A and two others for condition B. After alignment and differential expression analysis using DESeq package, I have a whole list of genes with fold changes of A vs B. Now, mu question is: where do I put a cutoff?

1. From a biological point of view, I'm tempted (as others have done the same) to set a FoldChage of 2 as a cutoff. 2 times more transcripts is somewhat significant at biological level for a cell. But is it really? If we assume it is, it brings me to the next point:
2. What is a cutoff for p-value? I'm tempted to use padj (hence FDR-corrected) and the hits I'll get are almost surely genuine (in fact, I tested those by qPCR and indeed they are differentially expressed from A vs B). However, am-I missing potentially interesting hits by being too much restrictive? Then, where do I set my cutoff?

FYI: I'm dealing with Illumina, single strand 50pb, non strand-specific, bacterial RNA-seq data.

Thank you all for your input on this,

TP

rnaseqrpkmdeseq

ADD COMMENT • link •

Not following

modified 8 months ago by seidel ♦ **4.6k** • written 8 months ago by ThePresident • **40**

2

8 months ago by
seidel ♦ **4.6k**
United States

I'll just echo what dpryan70 said in a comment, where you set your cutoffs depends completely on what you plan to do with the results. If you have an assay to easily screen through lots of genes, then you can be liberal about your cutoff, whereas if follow up involves heavy investment then you would be much more stringent. You might also use different cutoffs for different purposes. For instance, a cutoff to select genes for qPCR validation may be different than a cutoff you would use for GO enrichment analysis.

In my experience, the magnitude of the numbers (fold change, p or q value) do not have any absolute meaning - i.e. an x-fold threshold that determines biological significance. Every data set is different, experimental systems are different, and I have to adjust both fold change and p-value restrictions on an experiment by experiment basis. It's often tempting to take the interpretations of false discovery rates associated with p-values literally, and easy to forget that the numbers are based on assumptions about distributions. The "true" and "false" used to describe positives and negatives are based on an ideal, and what is actually true and false are difficult to know. There's also the issue of conflating significance with importance (avoid "the cult of statistical significance"). Many people adjust p-values and have nothing "significant" left, yet there is plenty of evident biology in the data staring them in the face. So pick some values that seem reasonable based on what you'd like to do with the results, and prepare to iteratively adjust your choices based on your needs.

ADD COMMENT • linkwritten 8 months ago by seidel ♦ **4.6k**

Your comment about X-fold thresholds is quite important. Even in a world with a perfect correspondence between RNA and protein level changes, a 10% change in one protein can be much more important than a 300% change in another. All the statistics in the world can't replace putting the data in a biological context.

ADD REPLY • linkwritten 8 months ago by Devon Ryan ♦ **11k**

Thank you guys again. It means a lot to have one's idea on all this. We often have our noses stick too close in our data that we lose the big picture. But, overall, that's exactly what I want to avoid: use "common" statistics to delimit my list of DE genes. I want to use p and q values along with biological reasoning behind it. The only problem is that journals often want you to use those parameters blindly. They want the $p < 0.05$ regardless of anything else (unfortunately, so does my adviser).Anyway, thank you again, it helped me in taking those analysis on another level and

defend it in front of those that believe that p-value is the top of the rock. PS - sorry for my bad English ;)

ADD REPLY • linkwritten 8 months ago by ThePresident • 40

1

8 months ago by
Devon Ryan ♦ 11k
Bonn, Germany

I generally filter by adjusted p-value (0.10 is a common threshold for adjusted p-values) and then rank by fold-change. You'll lose real and meaningful changes regardless of what you do, so don't fixate too much on that.

ADD COMMENT • linkwritten 8 months ago by Devon Ryan ♦ 11k

Thank you for your answer. I agree with you, we need to cut somewhere and we'll lose meaningful data regardless of cutoff... that's why we set limits like pval < 0.05 ou padj < 0.1. It's just that I'm not statistician, so I have no clue how much it really means to set a cutoff for padj at 0.1. Is that threshold low, medium, high? And I hate to use something just because it's common practice... however I don't have enough statistical knowledge to accurately judge by myself! ;)

ADD REPLY • linkwritten 8 months ago by ThePresident • 40

1

Well, high, medium and low are subjective terms, so you'll never get an answer to that. In general, that's probably a medium threshold for general use. The most appropriate threshold will depend on what you want to do with the results. If you're going to do something expensive and time consuming, like making a bunch of transgenic mice or designing a drug trial, then you'll want a higher threshold. Generally, people will do various validations, so that'll give you a better idea if perhaps you might benefit from changing the threshold.

ADD REPLY • linkwritten 8 months ago by D