t-tests, many genes

Kasper D. Hansen

Fall 2022

Many genes

In genomics, we have a large number of genes.

This imposes a **burden** (also called a multiple testing burden); we can think of differential expression as a “fishing expedition”.

But it also poses an opportunity. With many genes, we have tools for assessing our assumptions.

We will now describe some of the problems and opportunities.

Introduction

We will consider a large amount of tests, indexed by g (one for each gene).

We have m such tests.

For each g we have a test statistic t_g and p-value p_g.

Assumption: we will assume the tests are **calibrated**, which means $P(\text{error} \mid H_0) = \alpha$ when we use a level of α. Or said differently, p_g is uniformly distributed if the null hypothesis is true.

Remember, the state of the world, 1 test

Decision	True DE (H_A)	True non-DE (H_0)
Significant	correct	error, type I
Not significant	error, type II	correct

- If the test is calibrated, we have $P(\text{error} \mid H_0) = \alpha$ when we use a level of α
- Power is $P(\text{correct} \mid H_A)$

A good test has high power while controlling the type I error rate.

The state of the world, many tests

Decision	True DE	True non-DE	Total
Significant	S	V	R
Not significant	T	U	m-R
Total	$m - m_0$	m_0	m

- S: True positives
- T: False negatives (type II errors)
- V: False positives (type I errors)
- T: True negatives
- m is given
- R (the number of tests we reject) is known.
- All other quantities are unknown.

Observation

If we have α chance of an error, we expect this many type I errors (m_0 is unknowable).

$$E(V \geq 1 \mid H_0) = m_0\alpha \leq m\alpha$$

If we test $10,000$ genes and most of them are not differentially expressed, we expect something like 500 false positives if we use $\alpha = 0.05$. That's bad!

Error rates

To control the number of errors, we first need to choose an error rate:

- FWER (Family-wise error rate): control $P(V \geq 1)$
- FDR (False discovery rate): control $E(V/\min(R,1))$, the expected number of false discoveries among the R discoveries we make.

Bonferroni correction

The Bonferroni correction attempts to control the FWER. If we set the individual tests level to be α_0, we're looking to satisfy

$$P(V = 0) \leq \sum_{g=1}^{m} \alpha_0 = m\alpha_0 \leq \alpha$$

leading us to set $\alpha_0 = \alpha/m$. We're using Bonferroni's inequality here.

This always control the FWER under very weak assumptions. It is known to be conservative.

Alternatives are the Holm procedure, which is uniformly more powerful than Bonferrroni (ie. always better).

Bonferroni is widely used in genetics.

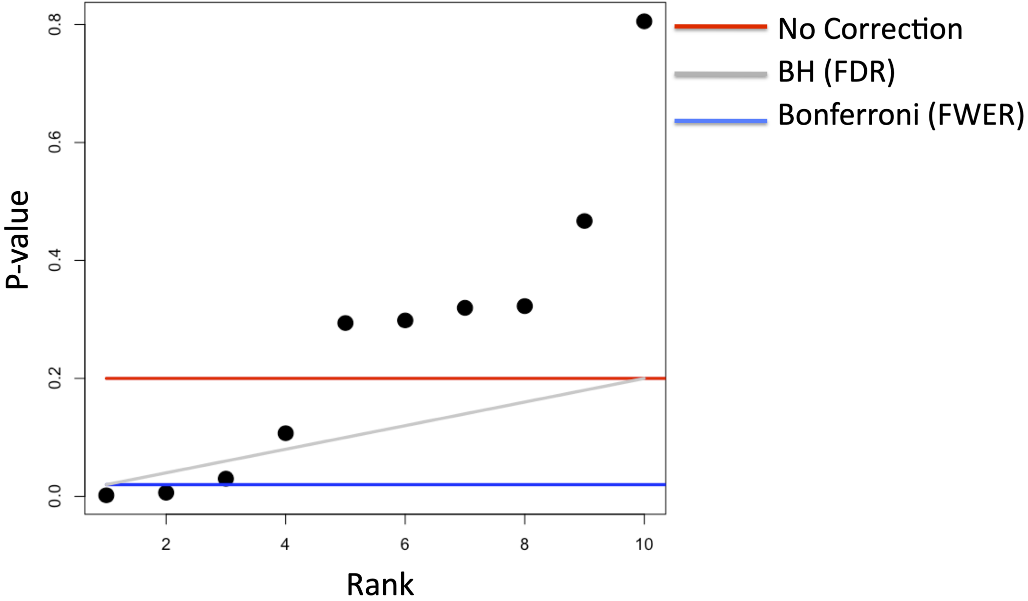
Benjamini-Horchberg correction

The Benjamini-Horchberg correction attempts to control the FDR.

1. Order the p-values $p_{(1)} \leq \cdots \leq p_{(m)}$
2. If $p_{(i)} \leq \alpha i/m$ then it is significant.

This depends on the sequence of p-values.

Illustration with 10 p-values, $\alpha = 0.2$



Opportunities

So far, we have heard the negatives about multiple tests: our chance of making errors increases substantially.

Now, let's discuss a few opportunities!

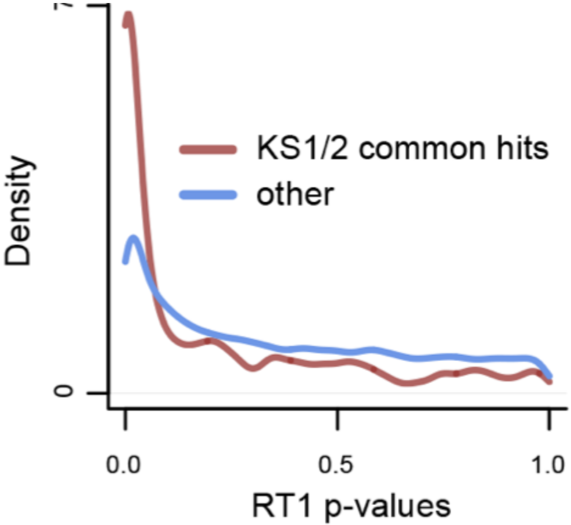
The statement that $P(\text{type I error}) = \alpha$ for a calibrated test can be re-stated as: under the null hypothesis, p-values are uniformly distributed.

Our observed p-value distribution is therefore a mixture of

1. a uniform distribution (most of the tests)
2. a distribution which should be concentrated around 0 (loosely, hard to make precise)

If we don't see this, we have a problem.

P-value distribution



From Luperchio, Boukas et al (eLife 2021)

QQ-plots

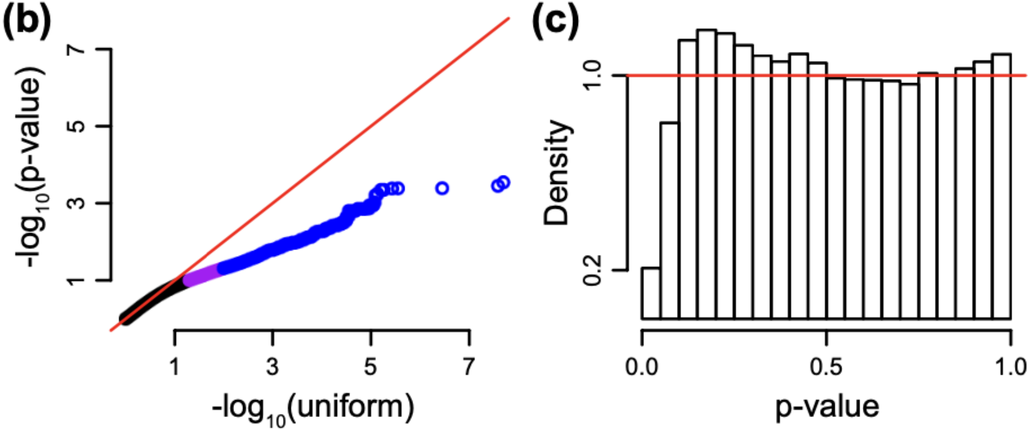
An alternative to histograms is to make a quantile-quantile plot. This (usually) requires that you're testing against a specific parameteric distribution.

You plot the quantiles of the observed statistic against the theroretical quantiles.

A QQ-plot should be a straight line around $y = x$. But equally important, if it is a straight line with a different slope/intercept it indicates that the two distributions are related by a scale-location transformation.

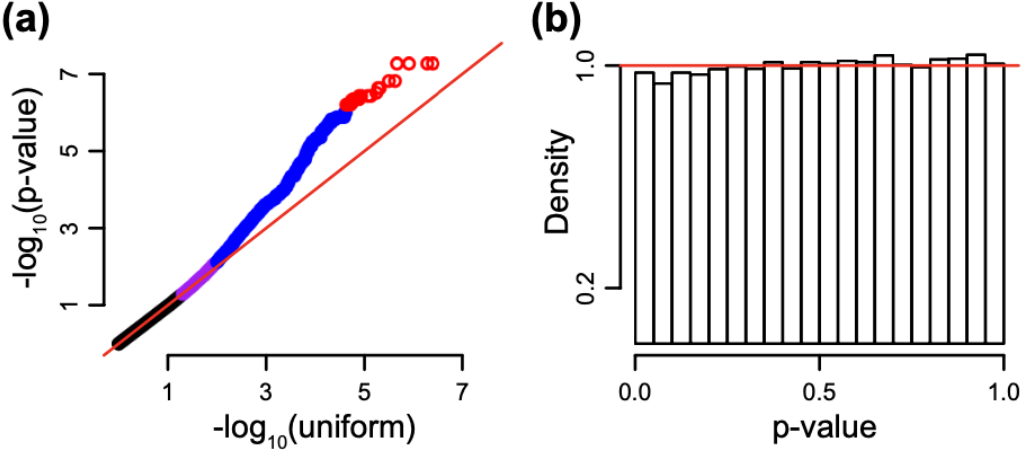
In some fields, there are immediate hypotheses that spring to mind when you see a QQ-plot which is a straight line but not at $y = x$. In popolation genetics, one possibility is population structure.

P-values, cont'd



Bad! From Fletez-Brandt et al (2020 bioRxiv)

P-values, cont'd



Corrected! From Fletez-Brandt et al (2020 bioRxiv)

Filtering

Remember we usually do some filtering to get from (say) 20,000 human genes to 8,000-12,000 expressed. Sometimes additional filtering is used.

Is this cheating? We decrease the number of tests by looking at the data. Because we have fewer tests, the Bonferroni correction gets less extreme.

This was discussed in Bourgon et al (PNAS 2010), which showed that filtering is permissible as long as you filter on a statistic which is independent from the test statistic. This is a complicated criteria which **depends on the test being used**.

Filtering on variance or mean is permissible for a t-test.

Testing and ranking

Let us take a different look at what is happening. We have sorted p-values:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$$

We're really trying to find a cutoff, below which we call the tests significant

$$\underbrace{p_{(1)} \leq \cdots \leq p_{(k)}}_{\text{significant}} \leq \underbrace{p_{(k+1)} \leq \cdots \leq p_{(m)}}_{\text{not significant}}$$

We have some critical value c such that $p_{(k)} \leq c < p_{(k+1)}$. Bonferroni says: choose c to be α/m.

All these methods (Bonferroni, Holm, Benjamini Horchberg) does not change the order of the p-values, they merely change c.

(In contrast to empirical Bayes variance shrinkage which changes the t-statistics and therefore the ranking)

For some questions, it is important to know the full extent of which genes are differentially expressed.

For some questions, ranking is good enough.

t-tests, variance, many genes

Under the assumption of equal variance (together with the minimal assumptions), we have

$$\mathrm{sd}(D_g) = \sigma_g \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where σ_g is the standard deviation of gene g.

It is hard to estimate standard deviations (variances) with small sample sizes, and in the t-test, this quantity is in the denominator: a potential nproblem!

An alternative model – which is clearly biologically unrealistic – is to assume that all genes have the same common variance. In this case, we have a large amount of data to estimate this parameter. This is an example of bias-variance trade-of:

Assumption	Consequence
gene specific variance	low bias, high variance
common variance	high bias, low variance

t-tests, variance shrinkage

Consider a quantity like

$$\hat{\sigma}^2_{\text{shrink}} = \frac{d_0 s^2 + d_g s_g^2}{d_0 + d_g}$$

Here, we shrink the estimates of the gene-specific variances towards a common variance, trying to achieve a balance between bias and variance.

How do we pick d_0, d_g? One approach to this is using Empirical Bayes on the variance.

Using these shrunken variances gives us a **moderated** t-statistic, which has been **extremely** useful in genomics. We are "borrowing information across genes". In genomics we tend to do this **only** for the variances; one could imagine doing something along these lines for the mean parameters as well.