The multiple testing problem in high-throughput biology

La Trobe University.
(Contact–Email: h.nguyen5@latrobe.edu.au; Website: hiendn.github.io)

Lectures in Mathematical Biology, 2020

Outline

- Hypothesis testing in high-throughput biology
- The multiple testing problem
- False discovery rate control
- The empirical-Bayes approach
- Genomics example

High-throughput biology

- Modern biological experiments acquire information regarding hundreds–millions of variables/features of interest about each observational unit, simultaneously.
- Each variable of interest requires assessment regarding its biological and statistical significance, via some technical mechanism.
- The inflated number of variables, combined with the relatively small sample sizes, increases the probability of false positives.

Examples of high-throughput experiments

- **Microarray/RNA-seq experiments** assess thousands–hundreds of thousands of different gene expressions.
- **Proteomic mass spectrometry experiments** can identify the expression of thousands–tens of thousands of different proteins.
- **MRI/fMRI experiments** can assess the effect of variables on millions of image voxels.
- **Geospatial observational studies** assess hundreds of thousands–millions of spatial image pixels for change or effects.

MRI example

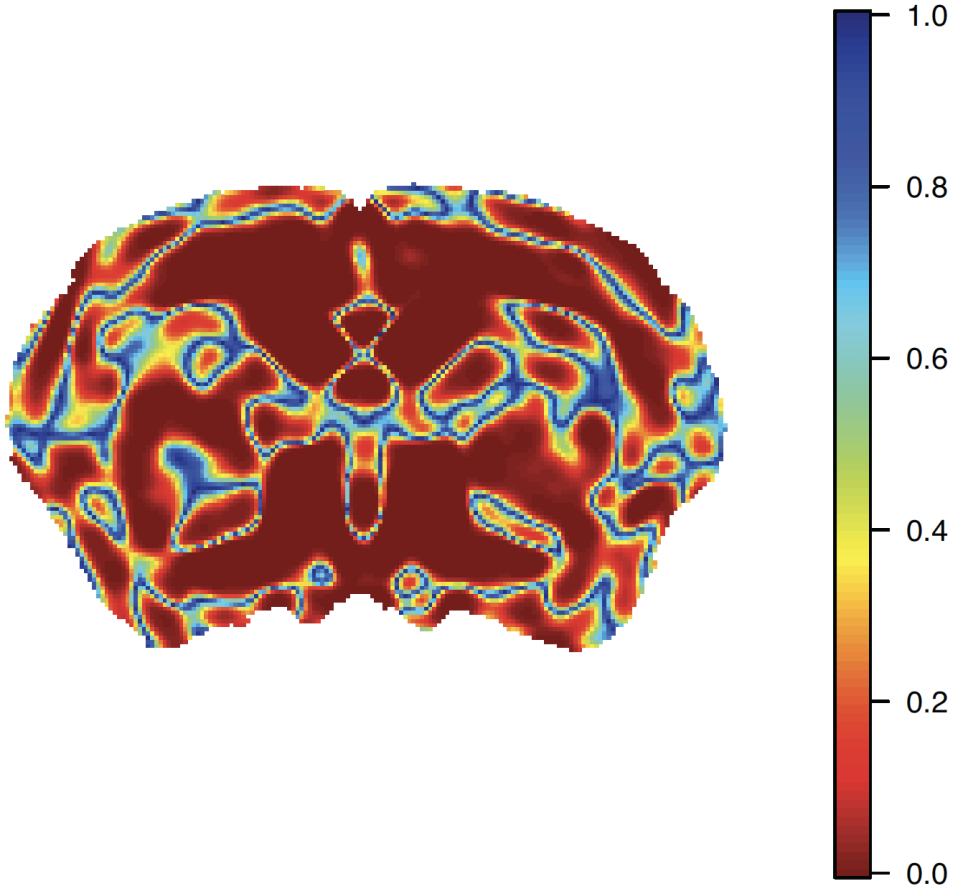

Figure: Coronal slice of a mouse brain MRI. Each pixel is a p-value. The brain volume is 2.8 million voxels (Nguyen et al., 2019).

Hypothesis testing

- Let H be a random **null hypothesis** that can either be true ($H = 0$) or false ($H = 1$).
- When testing H, we compute a ***p*-value** $P \in [0, 1]$ (summarizing our evidence) and compare it against the **level of significance** $\alpha \in [0, 1]$, where we assume that

$$\Pr(P \leq \alpha | H = 0) \leq \alpha.$$

- Let $r_\alpha(P)$ be the r**ejection rule** of H, where

$$r_\alpha(P) = \begin{cases} 0 \text{ (do not reject } H) & \text{if } P > \alpha, \\ 1 \text{ (reject } H) & \text{if } P \leq \alpha; \end{cases}$$

then we are accepting a **Type-I error rate** of:

$$\Pr(r_\alpha(P) = 1 | H = 0) \leq \alpha.$$

The multiple testing problem

- If we set $\alpha = 0.05$ then we accept that the probability of making a Type-I is at most $1/20$.
- Now suppose that we test n independent hypotheses $H_1, \ldots, H_n$, simultaneously, using the rejection rule r_α applied on their p-values $P_1, \ldots, P_n$; then the fact above implies

$$\mathrm{E}\left[\text{Total number of Type-I errors}\right] \leq \alpha n = n/20.$$

- Suppose that $n = 10000$ (modest modern scenario); then

$$\mathrm{E}\left[\text{Total number of Type-I errors}\right] \leq \frac{10000}{20} = 500.$$

The famous (infamous) salmon

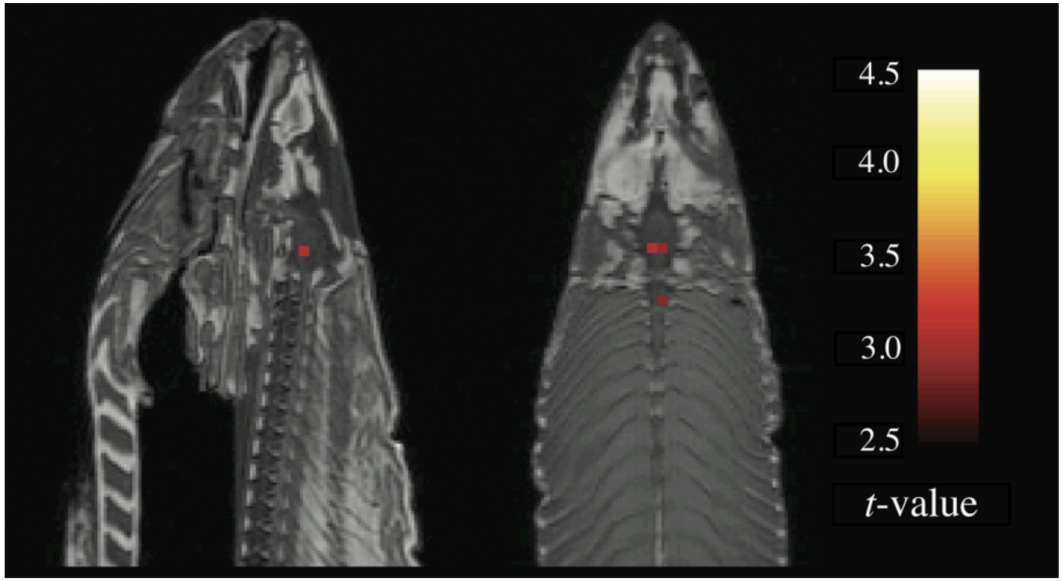

Figure: Significance map from an fMRI study of a dead salmon (Bennett et al., 2009).

The false discovery rate

- The rejection rule r_α is an individual test-focused rule and should not be applied to a large sample of hypotheses.
- We should subject the tests of the sample $H_1, \ldots, H_n$ to criteria that adequately accounts for the sampling property of the hypotheses.
- Let R be the **total number of rejected hypotheses** $H_1, \ldots, H_n$ according to some rule, and let V be the **number of these hypotheses that are falsely rejected**; then as famously suggested by Benjamini and Hochberg (1995), we can control the **false discovery rate** (FDR):

$$\mathrm{FDR} = \mathrm{E}(V/R)\,(\text{assuming } R > 0),$$

instead of the Type-I error.

Importance of the FDR

Figure: Web of Science citation report for Benjamini and Hochberg (1995).

The Benjamini-Hochberg method

- Like controlling the Type-I error probability below some level $\alpha \in [0,1]$; we wish instead to control the FDR so that

$$\text{FDR} \leq \beta, \ \beta \in [0,1].$$

- Let

$$P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(n)}$$

be the **order statistics** of the p-values $P_1, \ldots, P_n$, and correspondingly arrange the hypotheses $H_1, \ldots, H_n$ in the same order: $H_{(1)}, \ldots, H_{(n)}$.

The Benjamini-Hochberg method

- Benjamini and Hochberg (1995) suggest that if $H_1, \ldots, H_n$ are **independent**, then we can control the FDR at the level β by rejecting only the hypotheses $H_{(1)}, \ldots, H_{(k)}$, where

$$k = \max\left\{ i \in \{1, \ldots, n\} \,|\, P_{(i)} \leq \frac{i}{n}\beta \right\}.$$

- Alternatively, if we define the rejection rule:

$$r_\beta^{\mathrm{BH}}(P_i) = \begin{cases} 0 \text{ (do not reject } H_i) & \text{if } P_i > \frac{k}{n}\beta, \\ 1 \text{ (reject } H_i) & \text{if } P_i \leq \frac{k}{n}\beta; \end{cases}$$

then

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R}\right) = \mathrm{E}\left[\frac{\sum_{i=1}^{k} [\![H_{(i)} = 0]\!]}{k}\right] \leq \beta.$$

When does the Benjamini-Hochberg method work?

Benjamini and Yekutieli (2001) proved that the rule r_β^{BH} correctly controls the FDR under the conditions:

(BH1) If $H_i = 0$, then

$$P_i \sim \mathrm{Uniform}\,[0,1]\,.$$

(BH2) The n hypotheses satisfies the so-called **positive regression dependence on subsets** assumption (implied by positive correlation between the p-values).

The Benjamini-Yekutieli method

- Benjamini and Yekutieli (2001) proved that we can drop (BH2) if we let

$$k = \max\left\{ i \in \{1, \ldots, n\} \,|\, P_{(i)} \leq \frac{i}{n} \left(\sum_{i=1}^{n} \frac{1}{j} \right)^{-1} \beta \right\}$$

  and reject the k hypotheses $H_{(1)}, \ldots, H_{(k)}$ using the rejection rule:

$$r_\beta^{\mathrm{BY}}(P_i) = \begin{cases} 0 \text{ (do not reject } H_i) & \text{if } P_i > \frac{k}{n} \left(\sum_{i=1}^{n} \frac{1}{j} \right)^{-1} \beta, \\ 1 \text{ (reject } H_i) & \text{if } P_i \leq \frac{k}{n} \left(\sum_{i=1}^{n} \frac{1}{j} \right)^{-1} \beta. \end{cases}$$

- Blanchard and Roquain (2008) show that r_β^{BY} is within an infinite family of FDR control rules that are all correct under (BH1).

When are p-values not uniform, under the null hypotheses?

Efron (2010) suggest the following examples:

- *The mathematical assumption for the hypothesis tests are not satisfied*. E.g., t-test p-values are calculated using the wrong degrees of freedom.
- *Models are misspecified*. E.g., p-values are obtained from regression models where some normalizing variables are missing.
- *Tests are conducted using correlated data*. E.g., the genetic material from a genomics experiment are obtained from related individuals.
- *Hypotheses under consideration are pre-filtered*. E.g., the researcher quality controls data by removing variables containing outliers.

The z-transformation

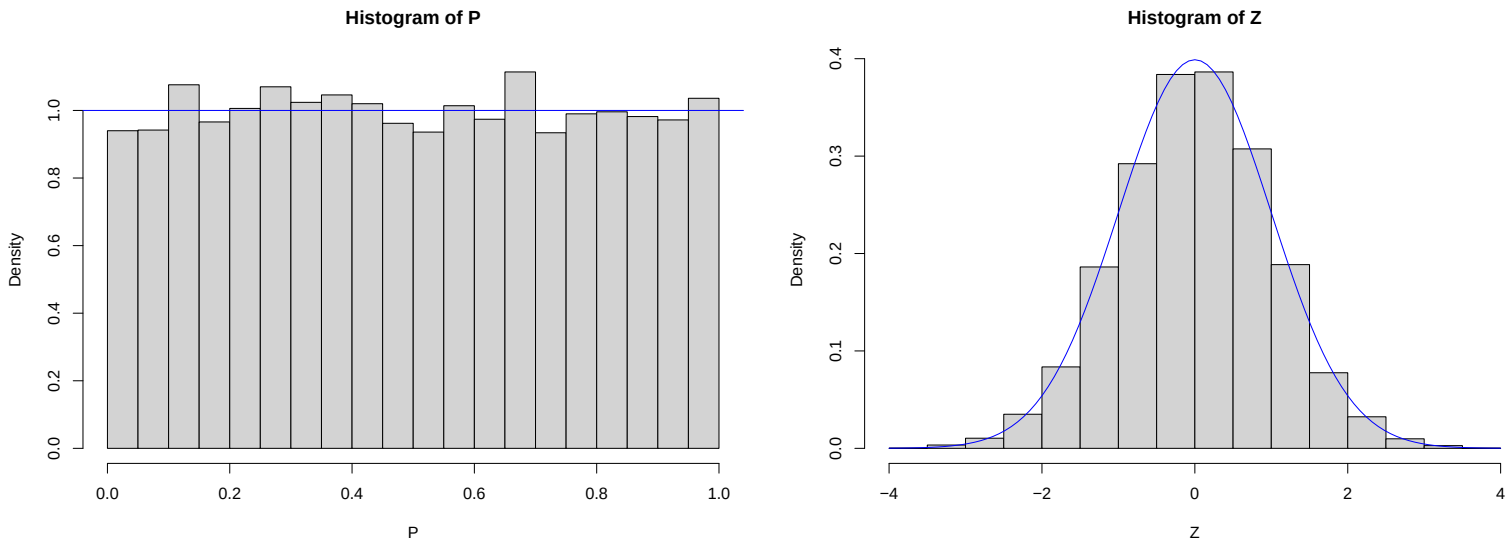

Figure: $P \sim \text{Uniform}\,[0,1]$, and $Z = \Phi^{-1}(1-P)$ and thus $Z \sim \text{N}(0,1)$.

Modeling the null distribution

- We observe that if P is not uniform, then the ***z*-score**

$$Z = \Phi^{-1}(1 - P)$$

will not be standard normal.

- If Z_i is the z-score of P_i and $H_i = 0$, then we can model the probability density function of Z_i by the normal PDF

$$f_0(z) = \phi\left(z; \mu_0, \sigma_0^2\right) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{1}{2}\frac{(z-\mu_0)^2}{\sigma_0^2}\right],$$

where $\mu_0 \approx 0$ and $\sigma_0^2 \approx 1$.

- Efron (2004) calls f_0 the **empirical null model** for the z-score.

Modeling non-null distribution

- If P_i is a p-value and $H_i = 1$, then on average, it should have a smaller value than when $H_i = 0$. (a well-ordered assumption, of sort).
- If Z_i is the z-score of P_i and $H_i = 1$, then we can model Z_i by the normal PDF

$$f_1(z) = \phi\left(z; \mu_1, \sigma_1^2\right),$$

where $\mu_1 > \mu_0$ (to enforce the well-ordering) and σ_1^2 is free to vary.

An empirical Bayes model for the z-score

- Suppose that each hypothesis H_i is equal to 0, with probability $\lambda_0 > 0$, and is equal to 1, with probability $\lambda_1 > 0$, so that

$$\lambda_0 + \lambda_1 = 1.$$

- Then, the empirical-Bayes model is the marginal PDF of Z_i:

$$f(z; \boldsymbol{\theta}) = \lambda_0 f_0(z) + \lambda_1 f_1(z) = \lambda_0 \phi\left(z; \mu_0, \sigma_0^2\right) + \lambda_1 \phi\left(z; \mu_1, \sigma_1^2\right),$$

  where $\boldsymbol{\theta}$ is a vector containing the parameter elements $\lambda_0, \lambda_1, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2$.
- If we knew the value of λ_0 and λ_1 then we can obtain the posterior distribution of H_i, conditioned on Z_i.
- Since we do not know any of the parameters, we must estimate $\boldsymbol{\theta}$ by the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, containing the elements $\hat{\lambda}_0, \hat{\lambda}_1, \hat{\mu}_0, \hat{\mu}_1, \hat{\sigma}_0^2, \hat{\sigma}_1^2$.

Estimating the FDR

- By Bayes' formula, we have the **estimated posterior probabilities**:

$$\hat{\tau}(Z_i) = \Pr_{\hat{\boldsymbol{\theta}}}(H_i = 0 | Z_i) = \hat{\lambda}_0 \phi\left(Z_i; \hat{\mu}_0, \hat{\sigma}_0^2\right) / f\left(Z_i; \hat{\boldsymbol{\theta}}\right).$$

- Suppose that we reject H_i using the rule:

$$r_c^{\mathrm{EB}}(Z_i) = \begin{cases} 0 \text{ (do not reject } H_i) & \text{if } \hat{\tau}(Z_i) > c, \\ 1 \text{ (reject } H_i) & \text{if } \hat{\tau}(Z_i) \leq c; \end{cases}$$

  where $c > 0$ is some cutoff value.

- We can estimate the FDR based on rule $r_c^{\mathrm{EB}}(Z_i)$ by

$$\overline{\mathrm{FDR}}(c) = \frac{\sum_{i=1}^n \hat{\tau}(Z_i)\, r_c^{\mathrm{EB}}(Z_i)}{\sum_{i=1}^n r_c^{\mathrm{EB}}(Z_i)},$$

  where $\mathrm{E}\left[\overline{\mathrm{FDR}}(c)\right] \to \mathrm{FDR}(c)$, as $n \to \infty$ (Nguyen et al., 2014).

Controling the FDR

- Using $\overline{\mathrm{FDR}}(c)$, we can model the FDR as a function of c.
- If we wish to control the FDR at the desired level $\beta \in [0,1]$, then we must find

$$\hat{c}_\beta = \sup\left\{c | \overline{\mathrm{FDR}}(c) \leq \beta\right\}.$$

- We can then apply the rejection rule:

$$r_\beta^{\mathrm{EB}}(Z_i) = \begin{cases} 0 \text{ (do not reject } H_i) & \text{if } \hat{\tau}(Z_i) > \hat{c}_\beta, \\ 1 \text{ (reject } H_i) & \text{if } \hat{\tau}(Z_i) \leq \hat{c}_\beta; \end{cases}$$

An example

- We analyze the `hivdata` data set from van't Wout et al. (2003) which contains $n = 7680$ normalized t-statistics corresponding to differential expression of genes related to HIV, computed from two-sample pooled t-tests from 4 HIV and 4 control patients.
- We follow the FDR control procedure that was considered in McLachlan et al. (2006).

References I

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188.

Bennett, C. M., Baird, A. A., Miller, M. B., and Wolford, G. L. (2009). Neuro correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. In *15th Annual meeting of the Organization for Human Brain Mapping*.

Blanchard, G. and Roquain, E. (2008). Tow simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992.

References II

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99:96–104.

Efron, B. (2010). *Large-scale Inference*. Cambridge University Press, Cambridge.

McLachlan, G. J., Bean, R. W., and Ben-Tovim Jones, L. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22:1608–1615.

Nguyen, H. D., McLachlan, G. J., Cherbuin, N., and Janke, A. L. (2014). False discovery rate control in magnetic resonance imaging studies via Markov random fields. *IEEE Transactions on Medical Imaging*, 33:1735–1748.

References III

Nguyen, H. D., Yee, Y., McLachlan, G. J., and Lerch, J. P. (2019). False discovery rate control for grouped or discretely supported p-values with applications to a neuroimaging study. *SORT*, 43:237–258.

van't Wout, A. B., Lehrman, G. K., Mikheeva, S. A., O'Keeffe, G. C., Katze, M. G., Bumgarner, R. E., Geiss, G. K., and Mullins, J. I. (2003). Cellular gene expression upon human immunodeficienct virus type 1 infection of CD4(+)-T-cell lines. *Journal of Virology*, 77:1392–1402.

Thank you!

Email: **h.nguyen5@latrobe.edu.au**

Website: **hiendn.github.io**