CS-E5875 High-Throughput Bioinformatics Introduction

Harri Lähdesmäki

Department of Computer Science
Aalto University

October 27, 2020

Contents

What is high-throughput bioinformatics?

- It is an interdisciplinary field that develops and applies methods for storing, retrieving, organizing and **analyzing** high-throughput biological data
- High-throughput technologies can be thought of as massively parallel automated methods to carry out a large number of individual experiments/biochemical tests simultaneously
- An example: a microarray or a sequencing machine can
  - Measure expression of tens of thousands of genes at once
  - Quantify genetic variants at millions of positions throughout a genome

  $\rightarrow$ Data are produced at a massive scale
- Suitable bioinformatics and statistical methods are needed to analyze and exploit these data
- Goals: too many to list here...

Data growth in genomics and bioinformatics

- Fast evolution in these fields – recent data explosion
- Consider for example:
  - When was the first genome sequence published?
  - When was the first version of the human reference genome sequence available?
  - How many human genomes have been sequenced by today?

History of genomics

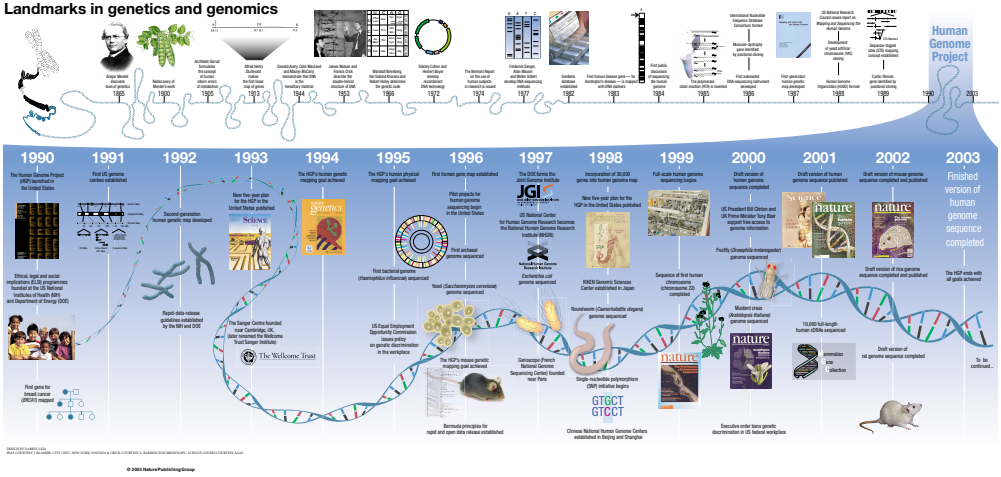


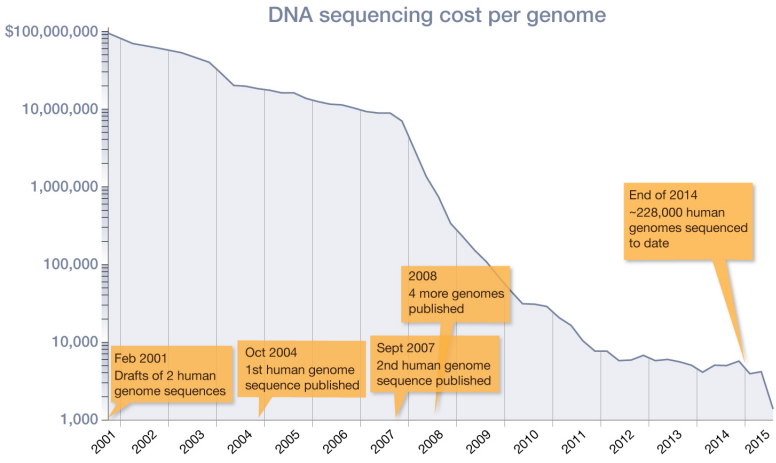
Figure from Nature Publishing Group

http://www.nature.com/nature/journal/v422/n6934/pdf/timeline_01626.pdf

Bioinformatics: historical perspective

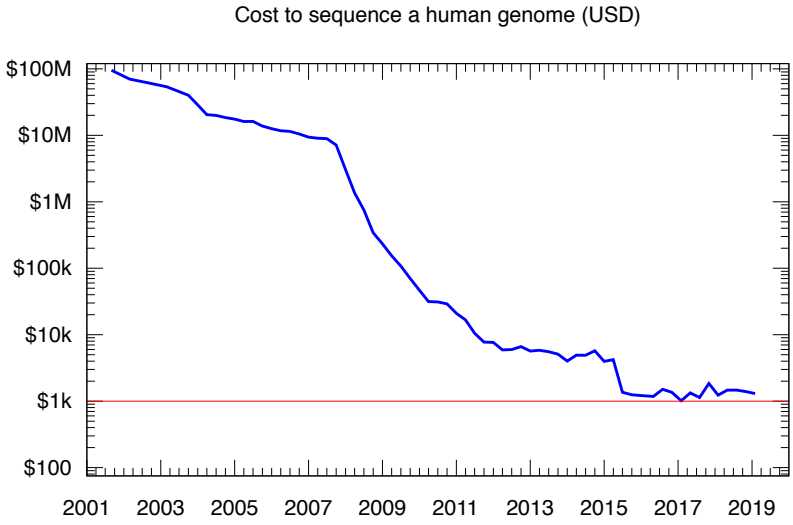
- 1956: The first protein sequenced / analysed
- 1965: The first atlas of protein sequences (printed book)
- 1970s: Term “bioinformatics” first used
- 1980s: Development of sequence alignment techniques
- 1980-90: Predicting RNA and protein structures
- 1990s: Prediction of genes
- 1990-2000s: Studies of complete genomes
- 2000+: Complete genomes, functional genomics, personalized medicine

Data growth: sequencing costs



http://learn.genetics.utah.edu/content/precision/time/

Data growth: sequencing costs



https://en.wikipedia.org/wiki/Whole_genome_sequencing

Data growth: no. of sequenced eukaryotic species

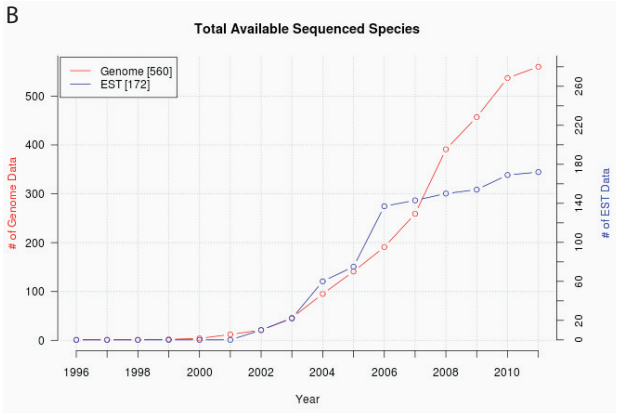


Figure from BMC Res Notes 4:338, 2011

- According to a Sanger Institute blog[1]: "There are fewer than 3,500 eukaryotic species with sequenced genomes. This represents less than 0.2 per cent of known eukaryotes."

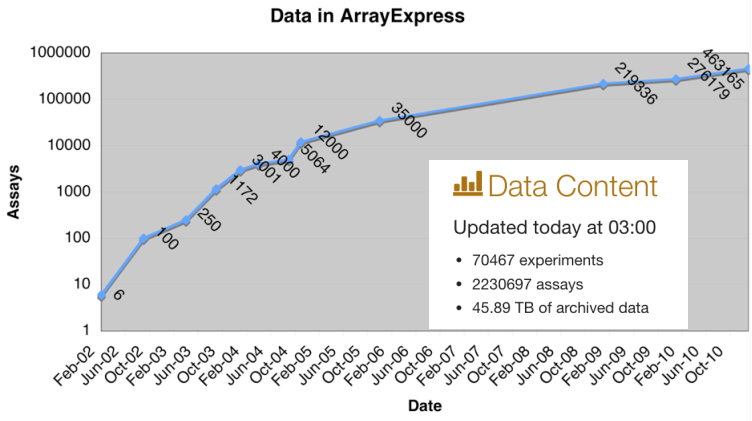
[1] `https://sangerinstitute.blog/2018/11/01/sequencing-all-life-on-earth-facts-and-figures/`

Beyond genome analysis

- After having sequenced the genome (e.g. human reference genome):
  - Characterize genetic variation between individuals
  - Identify the location of genes
  - Analyze gene functions, interactions, and regulation
  - Quantify and analyze epigenomics
  - Characterize dynamic properties of genome and functional genomics
  - Analyze genetics, functional genomics, epigenomics in the context of biomedicine
  - ...
  - Translate this data / knowledge for health and disease

Data growth: functional genomics assays in ArrayExpress

- ArrayExpress: a repository of functional genomics experiments, containing gene expression data from microarray and high-throughput sequencing experiments



More info: Nucl. Acids Res. 39 (suppl 1): D1002-4, 2011

Contents

Statistical hypothesis testing

- Hypothesis testing is a main inferential statistics concept that we will use throughout this course
- We will briefly review the basics of hypothesis testing
  - For this part, we follow closely parts of Jeremy Orloff's and Jonathan Bloom's excellent lecture notes material "Null Hypothesis Significance Testing" (Orloff and Bloom, 2014)
  - You may also refer to several / any statistics book
- Conceptually speaking, the so-called Newman-Pearson hypothesis testing framework asks if the observed data is outside the region where we expect the data to be
  - If it is, then we have evidence to reject our initial conservative expectation / hypothesis

Null hypothesis testing

- Key concepts:
  - H_0: the null hypothesis. This specifies the default assumptions for the model that generates the data
  - H_A: the alternative hypothesis (also denoted as H_1). We are interested in testing the null hypothesis; if null is rejected we accept the alternative hypothesis as the best explanation for the data
  - T: the test statistic, computed from the observed data
  - Null distribution: the probability density of the test statistic, assuming the null hypothesis holds true
- Typically the null hypothesis is chosen to be a simple or conservative hypothesis, which we reject if we have sufficient amount of evidence to reject H_0

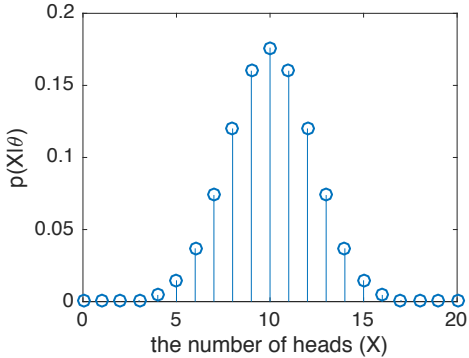
Example: coin flipping

- We flip a coin N times to test whether the coin is fair or unfair
- The rational is to check whether our coin results in unexpectedly few or many heads/tails
- Let θ denote the probability that the coin flipping results in a head (or tail), then:
  - Null hypothesis: $H_0 =$ "the coin is fair", i.e. $\theta = 0.5$
  - Alternative hypothesis: $H_A =$ "coin is not fair", i.e. $\theta \neq 0.5$
  - Test statistic: $T =$ number of heads in N flips
  - Null distribution: assuming the null hypothesis holds, the number of heads follows binomial distribution

$$T \sim \text{binomial}(N, 0.5)$$

Example: coin flipping

- The probabilities of obtaining any number of heads (between 0 and 20) from 20 coin flipping experiments are shown below (here X is used to denote the test statistic):



- So, is it "too unlikely" to observe e.g. as many as 15 heads? What about observing as few as 5 heads?

p-value

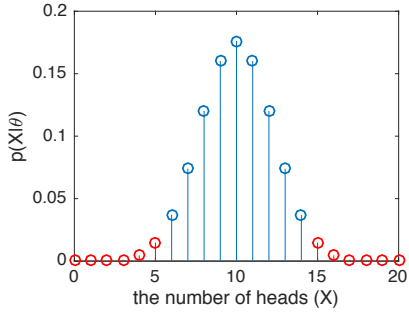
- For a given realization $T = t$, the p-value is the probability of seeing data / test statistic at least as extreme as t

$$p = P(\text{test statistic at least as extreme as } t)$$

- "At least as extreme as" depends on the hypothesis test / test statistic / experimental design
- Standard hypothesis tests are either one-sided or two-sided, i.e.,
  - One-sided: the test statistic can have significantly low values or high values (but not both)
    - One-sided test has directionality
  - Two-sided: the test statistic can have both significantly low values and high values

Example: coin flipping cont'd

- The coin flipping test is two-sided, because the number of heads can be either low or high
- The probability of obtaining T smaller than 6 or larger than 14 is $p \leq 0.05$
  - p-value of smaller than 0.05 is a commonly used threshold
  - The extreme values (red) form the *rejection region*
  - The typical values (blue) form the *"acceptance" region*
  - In the "acceptance" region we do not have enough evidence to reject H_0



Types of null hypothesis

- Simple hypothesis: a null hypothesis that specifies the population distribution exactly
  - E.g. data / test statistic is sampled from a given normal distribution with known mean and variance
- Composite hypothesis: a null hypothesis that does not specify the population distribution completely
  - E.g. data / test statistic is sampled from a given normal distribution with known mean but unknown variance
- Exact / point hypothesis: a null hypothesis that specifies an exact parameter value, e.g., $\text{mean} = 0$
- Inexact hypothesis: a null hypothesis that specifies a range of parameter values, e.g., $\text{mean} \leq 0$
- Our coin flipping example has a null hypothesis that is simple and exact

t-test

- In many applications data is assumed to be normally distributed
- Two-sample t-test can be applied to test the means of two samples which are assumed to be drawn from two normal distributions (with the same variance here)

$$\begin{aligned} x_1, \ldots, x_n &\sim N(\mu_1, \sigma^2) \\ y_1, \ldots, y_m &\sim N(\mu_2, \sigma^2) \end{aligned}$$

- Unknowns: μ_1, μ_2, and σ^2
- The null hypothesis H_0: $\mu_1 = \mu_2$
- The alternative hypothesis H_A: $\mu_1 \neq \mu_2$

t-test

- The test statistic T (T is the random variable, t is a particular realization of T)

$$t = \frac{\overline{x} - \overline{y}}{s},$$

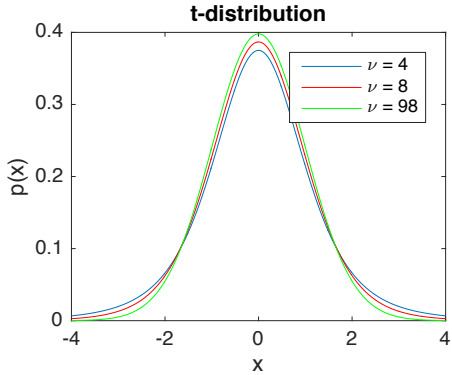
where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ and s^2 is the pooled variance

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right) \quad \text{and} \quad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- The null distribution: $p(T|H_0)$ can be shown to be the t-distribution with $n + m - 2$ degrees of freedom

t-test

- t-distribution for different degrees of freedom



t-test

- One-sided p-value (right side): $p = P(T > t|H_0)$
- One-sided p-value (left side): $p = P(T < t|H0)$
- Two-sided p-value: $p = P(|T| > |t|)$

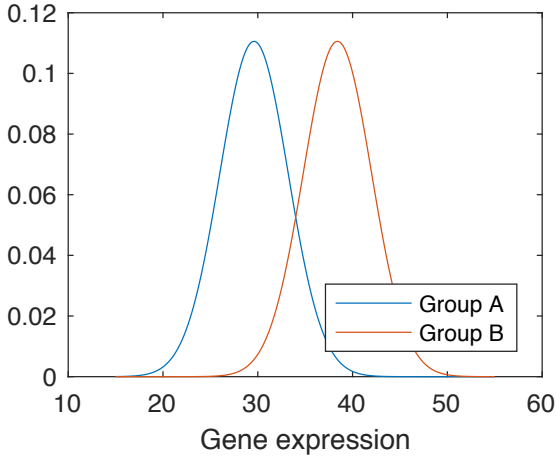
t-test

- An example: let us assume that we are interested in quantifying whether a gene of interest is differentially expressed between two groups A and B (say, between healthy and diseased individuals)
- Measured gene expression values are

$$\begin{aligned} \text{Group } A: &\quad 32, 25, 36, 27, 28 \\ \text{Group } B: &\quad 29, 48, 39, 37, 39 \end{aligned}$$

t-test

- We can explore the data & question by drawing estimated normal densities for both groups



t-test

- For quantitative inference, we can use the t-test
- The value of the t-statistic for our data is -2.4388
- In general, we may not know whether our gene can be up- or down-regulated and we need to apply two-sided test and obtain a p-value of 0.0406
- If we know that the expression value in group B can only be lower, we can apply one-sided test and obtain a p-value of 0.0203

Contents

Types of error

- Two types of errors can be made in a hypothesis testing
  - Type I error: null hypothesis H_0 is true but we reject that in favour of H_1. This incorrect decision results in a false positive.
  - Type II error: null hypothesis H_0 is not true but we do not reject H_0. This incorrect decision results in a false negative.

Table of error types		**Null hypothesis (H_0) is**	
		Valid/True	**Invalid/False**
Judgment of Null Hypothesis (H_0)	**Reject**	Type I error (False Positive)	Correct inference (True Positive)
	Accept	Correct inference (True Negative)	Type II error (False Negative)
Type-1 = True H_0 but reject it (False Positive) **Type-2 = False H_0 but accept it (False Negative)**			

Figure from (Wikipedia)

Power of a test

- Significance level of a test (often called α) is defined to be the probability that we incorrectly reject H_0

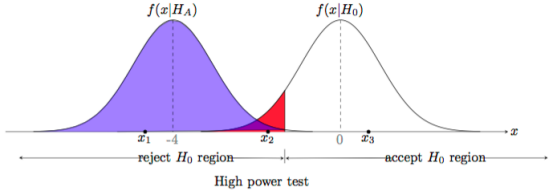
$$\text{Significance level} = P(\text{reject } H_0 | H_0) = P(\text{type I error})$$

- Power of a test is defined to be the probability that we correctly reject H_0

$$\begin{aligned} \text{Power} &= P(\text{reject } H_0 | H_A) \\ &= 1 - P(\text{do not reject } H_0 | H_A) \\ &= 1 - P(\text{type II error}) \end{aligned}$$

Power of a test

- Figure from (Orloff and Bloom, 2014) below illustrates the concept of power
  - Shaded area below $f(x|H_0)$ represents the significance
  - Shaded area below $f(x|H_A)$ represents the power: the probability that the test statistic is in the rejection region of H_0 when H_A is true
  - Note that the hypothesis testing works without knowing / caring about $f(x|H_A)$



Hypothesis test design

- Choose the null hypothesis H_0
- Decide if your alternative hypothesis is one-sided or two-sided
- Choose a test statistic
- Choose a significance level
- Determine the power (for different values of the alternative hypothesis)

Contents

Multiple testing

- Multiple testing problem occurs when a statistical analysis and decision making involves multiple simultaneous statistical hypothesis tests
- The p-values (i.e., confidence levels) described above are valid for a single test
- Consider the previous example of comparing gene expression (for gene x_1) between Groups A and B
  - If 5% confidence level is used for a single test, then there is only 0.05 probability that null hypothesis is rejected incorrectly
  - If the test is applied to 100 genes ($x_i, i \in \{1, \ldots 100\}$) for which the null hypothesis holds (i.e., they are not differentially expressed), then the expected number of genes for which the null hypothesis is rejected incorrectly is 5

Multiple testing

- Multiple testing problem occurs when a statistical analysis and decision making involves multiple simultaneous statistical hypothesis tests
- The p-values (i.e., confidence levels) described above are valid for a single test
- Consider the previous example of comparing gene expression (for gene x_1) between Groups A and B
  - If 5% confidence level is used for a single test, then there is only 0.05 probability that null hypothesis is rejected incorrectly
  - If the test is applied to 100 genes ($x_i, i \in \{1, \ldots 100\}$) for which the null hypothesis holds (i.e., they are not differentially expressed), then the expected number of genes for which the null hypothesis is rejected incorrectly is 5

$\rightarrow$ Hypothesis testing will lead to many false positives if the p-values are not corrected for multiple testing

- Multiple testing is a real issue in many (all?) bioinformatics applications
  - Differential gene expression analysis
  - Detecting disease associated genomic variant
  - Detection of protein binding sites along whole genome from ChIP-seq
  - ...

Multiple testing problem[2]

- Lets assume we have m independent hypothesis $H_0^{(1)}, \ldots, H_0^{(m)}$ and the null hypothesis holds for every one of them (that's a boring assumption to start with, but lets continue with that assumption anyways)
- If we make m independent tests anyway with significance level α, then each of the m tests will be significant with probability α
- Now the number of false positives X will have a distribution

$$X \sim \text{Binomial}(m, \alpha)$$

  (recall the coin flipping, now with a biased coin)
- The expectation of a binomial distribution is $E(X) = m\alpha$
- Once again, if we want to carry out a test e.g. for all approx. 20000 human genes, then the expected number of false positives (assuming null hypothesis holds for all) is $20000 \cdot 0.05 = 1000$

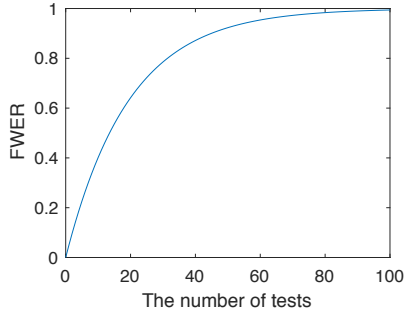
[2]From here onwards, parts of the slides follow Sections 7.2.2–7.2.4 from (Wilkinson, 2017). You can also check Section 18.7 from (Hastie et al., 2017)

Family-wise error rate

- Type I error
  - Null hypothesis H_0 is true but it is rejected in favour of H_1
- Assume m independent tests for which the null hypothesis is true, then the probability that any of the hypothesis will be rejected with significance level α is

$$\overline{\alpha} = 1 - (1 - \alpha)^m$$

  i.e., the probability of making one or more type I errors
- This is also called the family-wise error rate (FWER)



Bonferroni correction

- Let $H_0^{(1)}, \ldots, H_0^{(m)}$ be a collection of hypotheses and $p_1, \ldots, p_m$ the corresponding p-values
- Let $I_0 \subseteq \{1, \ldots, m\}$ be the subset of the $m_0 = |I_0| \leq m$ (unknown) true null hypotheses
- Bonferroni correction is defined as follows:
  - Given the original significance level α and the number of statistical tests m, then Bonferroni correction will reject only those null hypothesis i for which $p_i \leq \alpha/m$
  - Equivalently, the multiple testing corrected p-value for the i^{th} test is then $\min\{mp_i, 1\}$

Bonferroni correction

- Let $H_0^{(1)}, \ldots, H_0^{(m)}$ be a collection of hypotheses and $p_1, \ldots, p_m$ the corresponding p-values
- Let $I_0 \subseteq \{1, \ldots, m\}$ be the subset of the $m_0 = |I_0| \leq m$ (unknown) true null hypotheses
- Bonferroni correction is defined as follows:
  - Given the original significance level α and the number of statistical tests m, then Bonferroni correction will reject only those null hypothesis i for which $p_i \leq \alpha/m$
  - Equivalently, the multiple testing corrected p-value for the i^{th} test is then $\min\{mp_i, 1\}$
- For the Bonferroni correction $\mathrm{FWER} \leq \alpha$ because

$$\mathrm{FWER} = P\left(\bigcup_{i \in I_0} p_i \leq \frac{\alpha}{m}\right) \leq \sum_{i \in I_0} P\left(p_i \leq \frac{\alpha}{m}\right) = m_0 \frac{\alpha}{m} \leq= \alpha$$

- The Bonferroni correction is conservative

False discovery rate

- False discovery rate (FDR) is the proportion of false positives among all positives

$$\text{FDR} = \frac{\#\text{false positives}}{\#\text{false positives} + \#\text{true positives}}$$

- Formally FDR is defined as the expectation of the above quantity
- FDR of 0.05 means that 5% of the rejected null hypothesis are false
- However, on the other hand, FDR of 0.05 suggests that 95% of the rejected hypothesis are still true findings
- A small fraction of false positives are often accepted as long as majority of the results are true

False discovery rate

- Lets again assume that we have m tests with p-values $p_1, \ldots, p_m$
- We can order the p-values in increasing order $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$
- The choice of significance level is equivalent to deciding how many of the smallest p-values to consider significant
  - Lets denote that number (a positive integer) by ℓ
- Because a significance level α corresponds to a particular cutoff ℓ, we denote that by $\ell(\alpha)$, giving a list of significant p-values, $p_{(1)}, p_{(2)}, \ldots, p_{(\ell(\alpha))}$
  - A small α results in a short list (small ℓ)
  - A larger α results in a longer list (larger ℓ)
  - $\ell(\alpha)$ is monotonically increasing in α

False discovery rate

- Lets assume that the number of true positives (for which the null hypothesis does not hold) is small compared to the number of tests m
- Thus, similarly as above, the number of false positives is still approximatively distributed as $X \sim \text{Binomial}(m, \alpha)$
- Thus, the FDR is (assuming $\ell(\alpha) \geq X$)

$$\text{FDR} \approx \frac{X}{\ell(\alpha)} \quad \text{and} \quad E(\text{FDR}) \approx \frac{E(X)}{\ell(\alpha)} = \frac{m\alpha}{\ell(\alpha)}$$

False discovery rate

- Lets assume that the number of true positives (for which the null hypothesis does not hold) is small compared to the number of tests m
- Thus, similarly as above, the number of false positives is still approximatively distributed as $X \sim \text{Binomial}(m, \alpha)$
- Thus, the FDR is (assuming $\ell(\alpha) \geq X$)

$$\text{FDR} \approx \frac{X}{\ell(\alpha)} \quad \text{and} \quad E(\text{FDR}) \approx \frac{E(X)}{\ell(\alpha)} = \frac{m\alpha}{\ell(\alpha)}$$

- Generally we want to limit the fraction of false positive findings (i.e., FDR) by a value q, thus

$$\frac{m\alpha}{\ell(\alpha)} \leq q \quad \Leftrightarrow \quad \alpha \leq \frac{q\ell(\alpha)}{m}$$

- One needs to choose a small enough α so that the above inequality holds
  - This is little tricky because $\ell(\alpha)$ depends on α too

False discovery rate

- To solve the inequality on the previous page, assume we have inverted the function $\ell(\cdot) : [0, 1] \to \{1, \ldots, m\}$ as $\alpha(\cdot) : \{1, \ldots, m\} \to [0, 1]$
- We can write

$$\alpha(\ell) \leq \frac{q\ell}{m}$$

- Then notice that the p-value threshold that gives a list of length ℓ is $p_{(\ell)}$, thus we have

$$p_{(\ell)} \leq \frac{q\ell}{m}$$

- Now we just need to run through all possible values of ℓ, from 1 to m, in order to find the largest value of ℓ that satisfies the inequality and to find $p_{(\ell)}$

Benjamini-Hochberg correction

- The Benjamini-Hochberg (BH) step-up procedure is commonly used in bio applications
- Let q be given and $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ be the ordered (from smallest to largest) list of the m p-values, then the BH procedure works as follows
  1. Find the largest k such that $p_{(k)} \leq \frac{k}{m} q$
  2. Then reject all $H_{(i)}$ for $i = 1, \ldots, k$
- For BH, the probability of expected proportion of false positives $\leq q$
- The FDR value q_k for each test k can be obtained from mapping

$$\min \left\{ \frac{m}{k} p_{(k)}, 1 \right\}$$

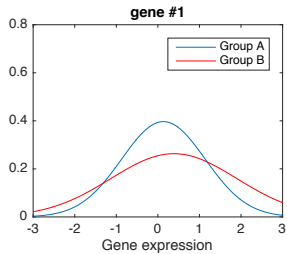
(and by guaranteeing that FDR values do not decrease as k increases)

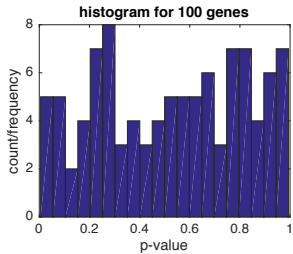
False discovery rate

- An example: Following the above example with one gene, let us now assume that we measure the expression of 100 genes for two groups, A and B. We have five replicate measurements (of 100 genes) from both groups.
- For each gene, expression values are normally distributed with means μ_A and μ_B and standard deviations $\sigma_A = \sigma_B$.

False discovery rate

- If $\mu_A = \mu_B = 0$ (and $\sigma_A = \sigma_B = 1$), the null hypothesis holds for all genes and in ideal case we should not detect any differentially expressed genes. However, the obtained p-values look as follows (histogram on right).

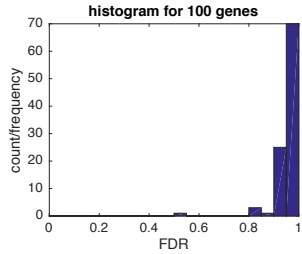




- We detect 5 genes with a p-value smaller than 0.05 (the magical threshold used in most of the fields of science)
  - Recall the definition of the significance level

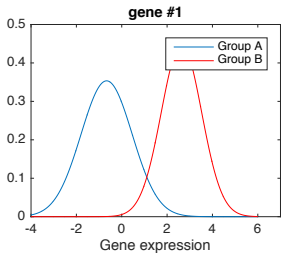
False discovery rate

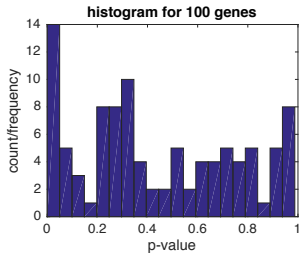
- If we correct the p-values for multiple testing using the Benjamini-Hochberg methods described above, we detect no genes that are statistically significantly differentially expressed.



False discovery rate

- Let us then see how FDR correction works if we have 10 truely differentially expressed genes and 90 non-differentially expressed genes with $\mu_A = 0$ and $\mu_B = 2$ for the differentially expressed genes.

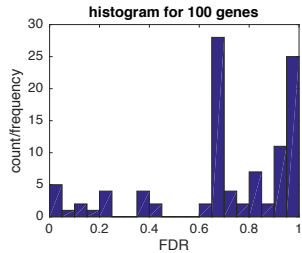




- We would now detect 14 genes with a p-value smaller than 0.05

False discovery rate

- If we correct the p-values for multiple testing using the Benjamini-Hochberg methods described above, we detect 5 genes that are statistically significantly differentially expressed.

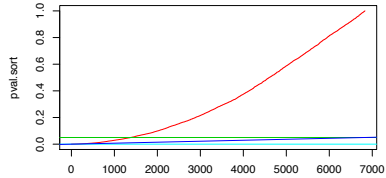


False discovery rate

- Consider an example from (Wilkinson, 2017): use t-test to identify genes differentially expressed in melanoma compared to healthy skin cells
- 6830 genes, i.e., $m = 6830$
- If we assumed that the null hypothesis holds for all genes, then the expected number of false positives would be $6830 \cdot 0.5 = 341.5$
- Using the nominal (non-corrected) p-values results in 1377 significantly differentially expressed genes, indicating that the data may contain a considerable number of truly differential genes
- The use of Bonferroni correction would give us only six genes that meet the stringent criterion of $p \leq 0.05/6830 \approx 0.0000073$
- BH correction method would give us 186 differentially expressed genes with a FDR threshold of 0.05

False discovery rate

- The figures below show
  - Ordered p-values (red)
  - The 0.05 uncorrected p-value cutoff (green)
  - The Bonferroni-corrected threshold (cyan)
  - The FDR threshold (dark blue)



Figures from (Wilkinson, YEAR)

References

- Hastie T, Tibshirani R, Friedman J, The Elements of Statistical Learning, Springer, 2009.
- Jeremy Orloff and Jonathan Bloom. "Null Hypothesis Significance Testing" I Class 17, 18.05, Spring 2014 (http://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading17b.pdf)
- Wilkinson DJ, Statistics for Big data Part 2: Multivariate Data Analysis using R (Lecture notes) available at https://www.staff.ncl.ac.uk/d.j.wilkinson/teaching/mas8381/notes14.pdf, November 19, 2017