Statistical Tests

Bioinformatica

Febrero-Junio 2020

Statistical Tests

POINTS OF SIGNIFICANCE

Significance, *P* values and *t*-tests

The *P* value reported by tests is a probabilistic significance, not a biological one.

Bench scientists often perform statistical tests to determine whether an observation is statistically significant. Many tests report the *P* value to measure the strength of the evidence that a result is not just a likely chance occurrence. To make informed judgments about

Statistical Tests

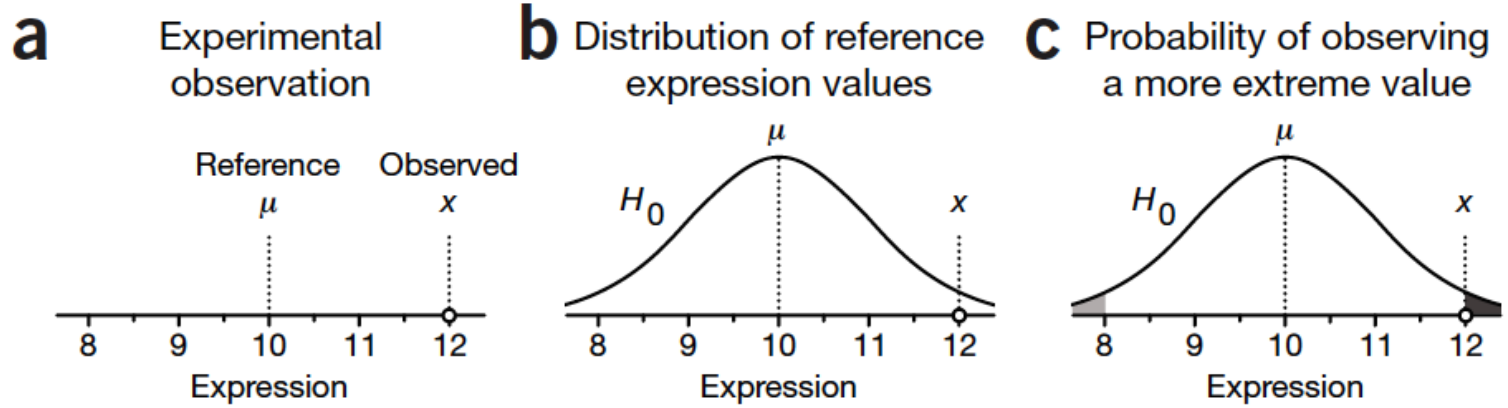


Figure 1 | The mechanism of statistical testing. (**a**–**c**) The significance of the difference between observed (x) and reference (μ) values (**a**) is calculated by assuming that observations are sampled from a distribution H_0 with mean μ (**b**). The statistical significance of the observation x is the probability of sampling a value from the distribution that is at least as far from the reference, given by the shaded areas under the distribution curve (**c**). This is the P value.

Statistical Tests

Unfortunately, the P value is often misinterpreted as the probability that the null hypothesis (H_0) is true. This mistake is called the 'prosecutor's fallacy', which appeals to our intuition and was so coined because of its frequent use in courtroom arguments. In the process of calculating the P value, we assumed that H_0 was true and that x was drawn from H_0. Thus, a small P value (for example, $P = 0.05$) merely tells us that an improbable event has occurred in the context of this assumption. The degree of improbability is evidence against H_0 and supports the alternative hypothesis that the sample actually comes from a population whose mean is different than μ. Statistical significance suggests but does not imply biological significance.

Statistical Tests

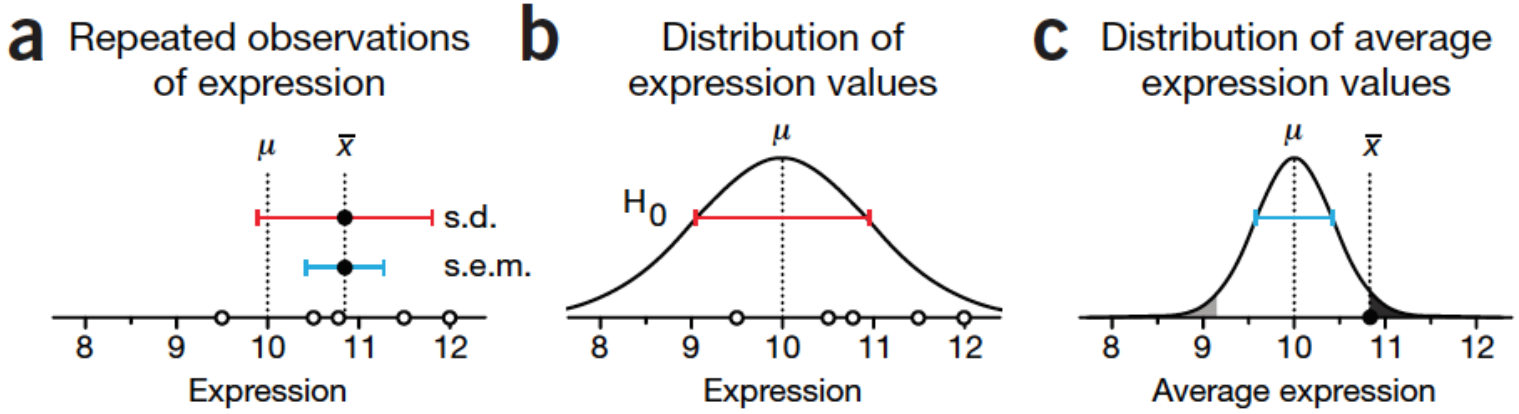


Figure 2 | Repeated independent observations are used to estimate the s.d. of the null distribution and derive a more robust *P* value. (**a**) A sample of $n = 5$ observations is taken and characterized by the mean $\bar{x}$, with error bars showing s.d. (s_x) and s.e.m. ($s_x/\sqrt{n}$). (**b**) The null distribution is assumed to be normal, and its s.d. is estimated by s_x. As in **Figure 1b**, the population mean is assumed to be μ. (**c**) The average expression is located on the sampling distribution of sample means, whose spread is estimated by the s.e.m. and whose mean is also μ. The *P* value of $\bar{x}$ is the shaded area under this curve.

T-distribution

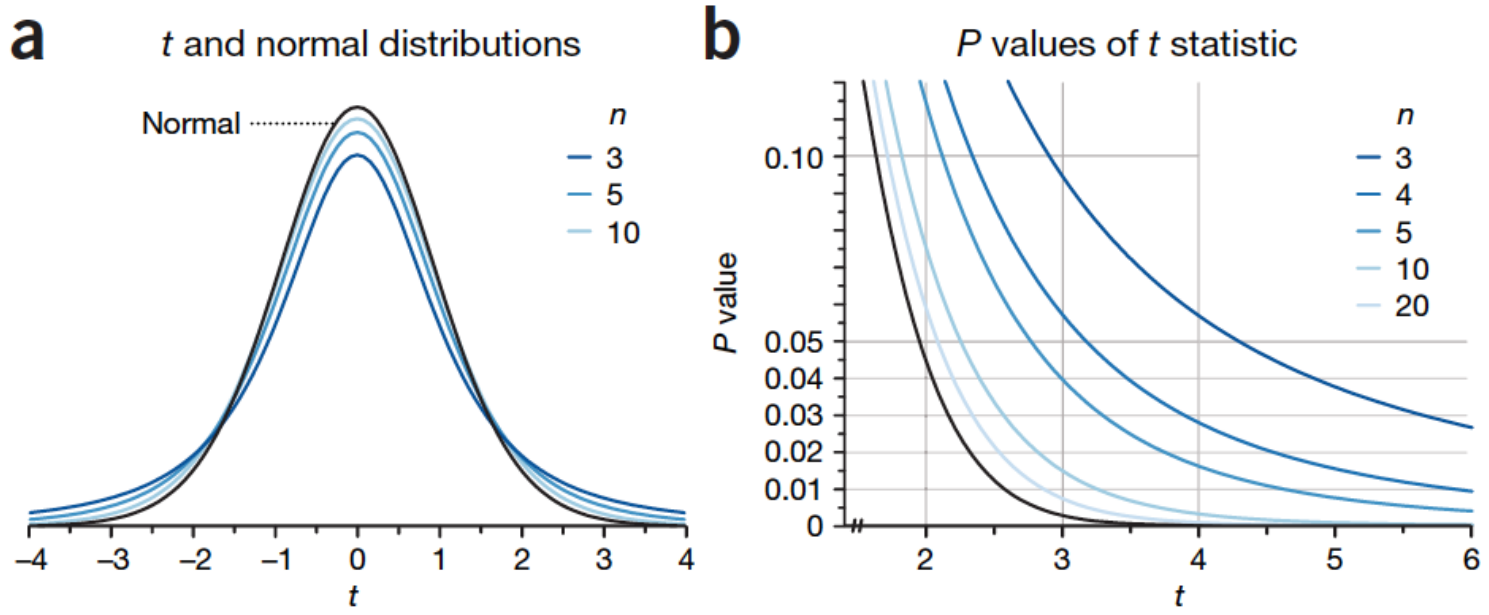


Figure 3 | The *t* and normal distributions. (**a**) The *t* distribution has higher tails that take into account that most samples will underestimate the variability in a population. The distribution is used to evaluate the significance of a *t* statistic derived from a sample of size *n* and is characterized by the degrees of freedom, d.f. = *n* – 1. (**b**) When *n* is small, *P* values derived from the *t* distribution vary greatly as *n* changes.

Statistical Tests

T-test

t statistic

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{s_X^2}{M} + \frac{s_Y^2}{N}}}$$

$\bar{Y}$ = mean sample Y
$\bar{X}$ = mean sample X
Sx = sd sample X
Sy = sd sample Y
M = length sample X
N = length sample Y

When M and N are large, this random variable is normally distributed with mean 0 and SD 1.
Assumes normal distribution.

R

t.test(rnorm(n=100,mean=4, sd=1), rnorm(n=100, mean=2, sd=1))$p.value

```
> t.test(rnorm(n=100,mean=4, sd=1), rnorm(n=100, mean=2, sd=1))$p.value
[1] 1.2842e-30
```

Wilcoxon or Mann-Whitney Test

U statistic

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

R1 = sum of ranks sample 1
n1 = length of sample 1
R2 = sum of ranks sample 2
n2 = length of sample 2

U = min(U1, U2)

$$z = \frac{U - m_U}{\sigma_U}$$

$$m_U = \frac{n_1 n_2}{2}$$

Does not assumes normal distribution.

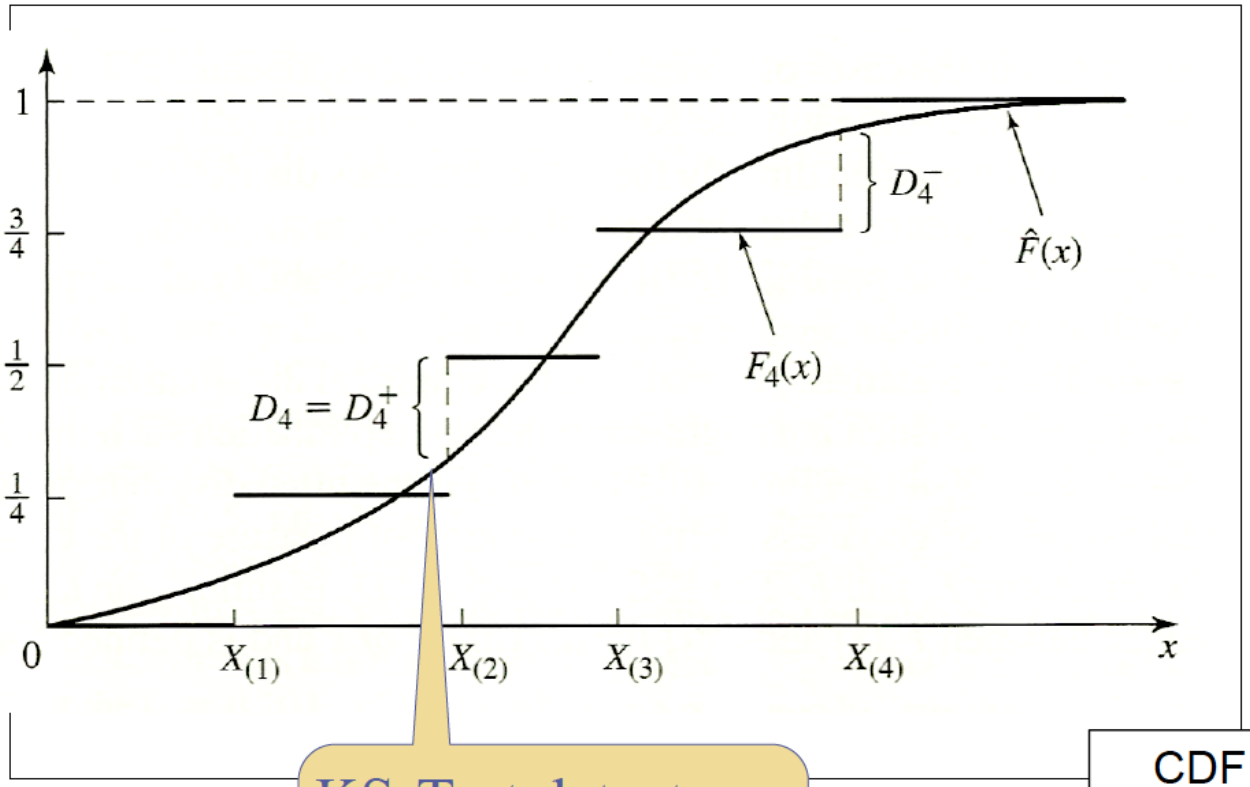
$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

R

wilcox.test(rnorm(n=100,mean=4, sd=1), rnorm(n=100, mean=2, sd=1))$p.value

```
> wilcox.test(rnorm(n=100,mean=4, sd=1), rnorm(n=100, mean=2, sd=1))$p.value
[1] 6.297855e-26
```

Kolmogorov-Smirnov Test



KS-Test detects the max difference

K-S test is useful when sample size is small

Test statistic

$D = max| F(x) - S_n(x)|$

CDF of the hypothesized distribution

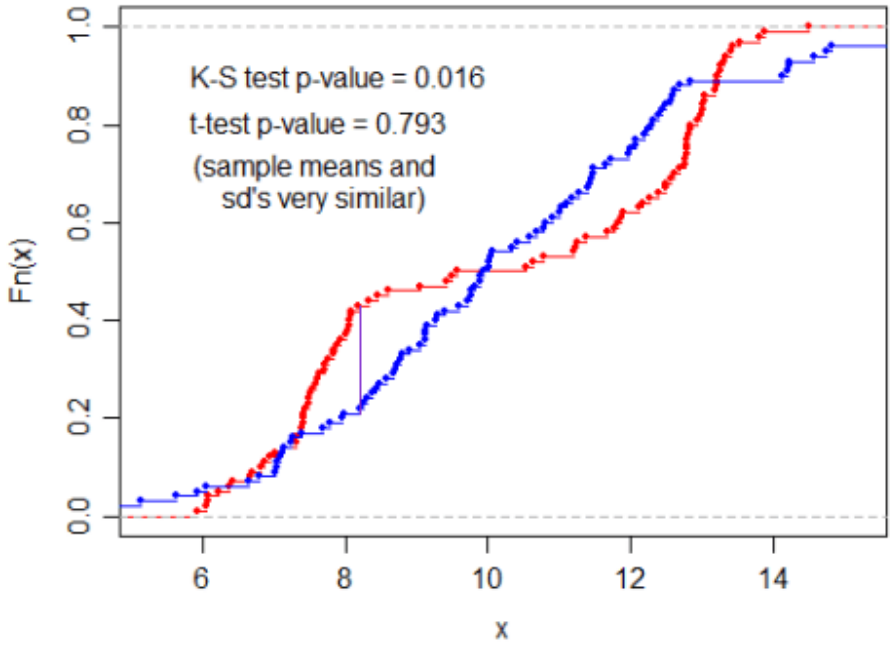
CDF of the empirical distribution constructed from the data

– If we have n observations $x_1, x_2, \ldots, x_n$, then

$S_n(x)$ = (number of $x_1, x_2, \ldots, x_n$ that are ≤ x) / n

Kolmogorov-Smirnov Test

K-S statistic



Does not assumes normal distribution.

R

ks.test(rnorm(n=100,mean=4, sd=1), rnorm(n=100, mean=2, sd=1))$p.value

```
> ks.test(rnorm(n=100,mean=4, sd=1), rnorm(n=100, mean=2, sd=1), alternative="l")$p.value
[1] 2.10494e-21
```

SAM

Significance analysis of microarrays applied to the ionizing radiation response

Virginia Goss Tusher*, Robert Tibshirani†, and Gilbert Chu*‡

3. For each gene, compute d-value (analogous to t-statistic).

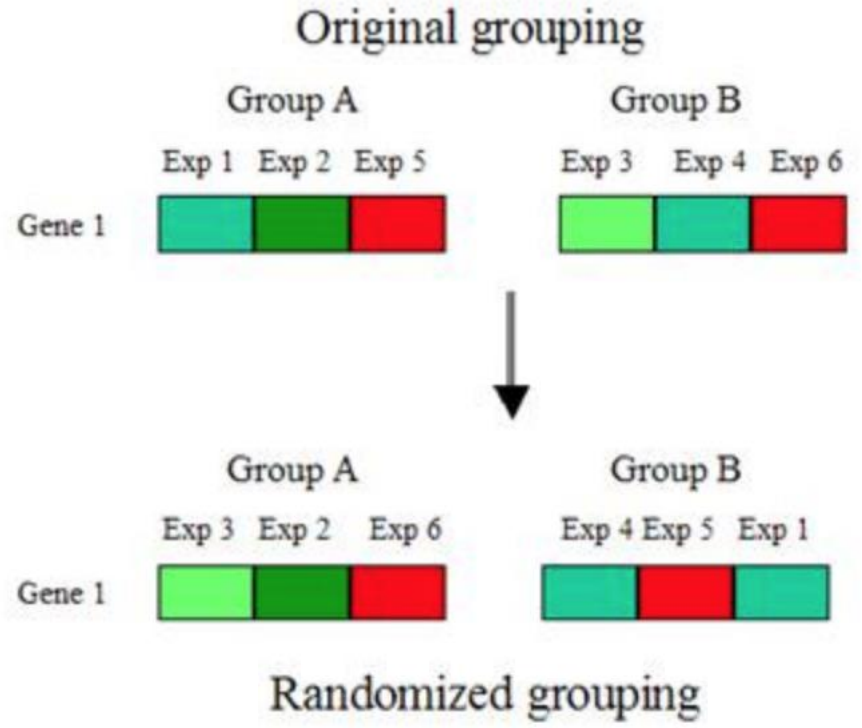
- d = Score
- s = Standard Deviation
- s_0 = Fudge Factor

$$r = \bar{x}_A - \bar{x}_B$$

$$d = \frac{r}{s + s_0}$$

$$s = \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \frac{\sum_{k=1}^{n_A} (x_k - \bar{x}_A)^2 + \sum_{k=n_A+1}^{n_A+n_B} (x_k - \bar{x}_B)^2}{n_A + n_B - 2}}$$

This is the **observed d-value**.

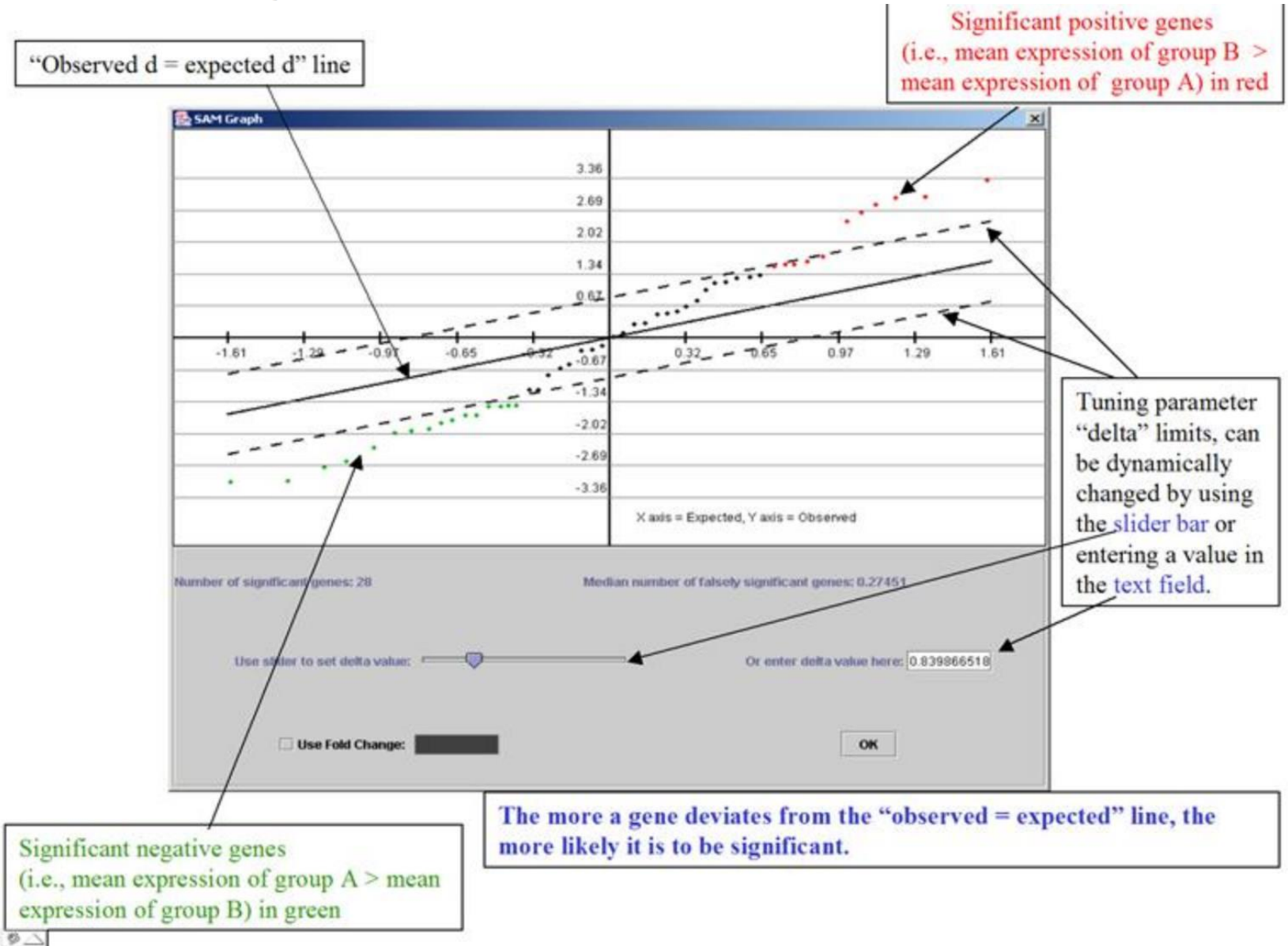


Permutation ⟶ **expected d-value**

SAM

Significance analysis of microarrays applied to the ionizing radiation response

Virginia Goss Tusher*, Robert Tibshirani†, and Gilbert Chu*‡



SAM

Significance analysis of microarrays applied to the ionizing radiation response

Virginia Goss Tusher*, Robert Tibshirani†, and Gilbert Chu*‡

Parameter	Number falsely significant	Number called significant	FDR
SAM			
$\Delta = 0.4$	134.9	288	47%
$\Delta = 0.5$	78.1	192	41%
$\Delta = 0.6$	56.1	162	35%
$\Delta = 0.9$	19.1	80	24%
$\Delta = 1.2$	8.4	46	18%

Normalization

Quantile Normalization

quantro: a data-driven approach to guide the choice of an appropriate normalization method

Stephanie C. Hicks and Rafael A. Irizarry

Genome Biology 2015 16:117

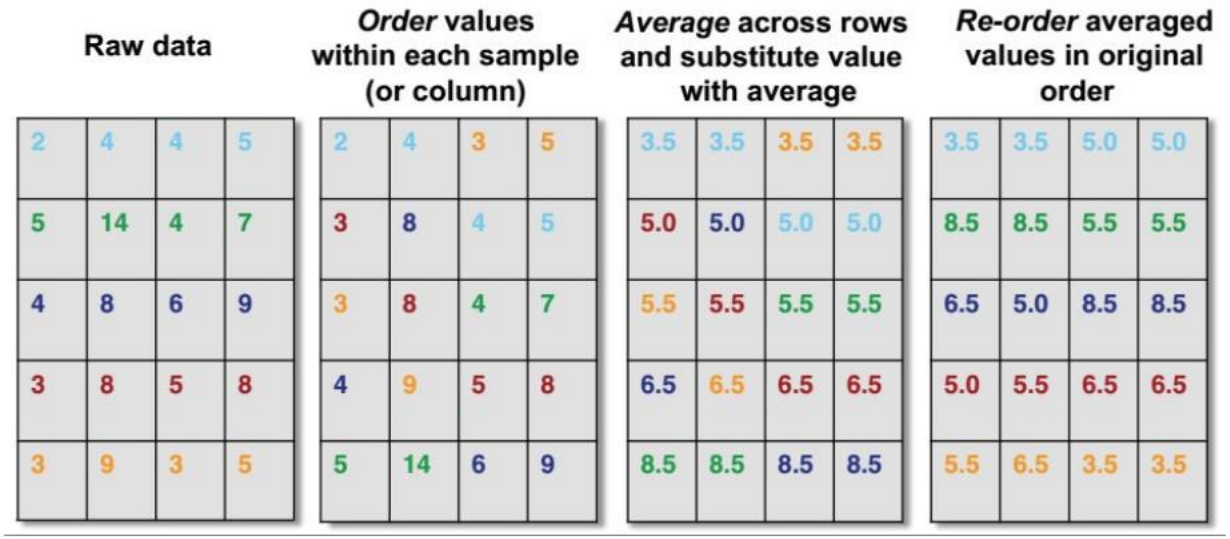
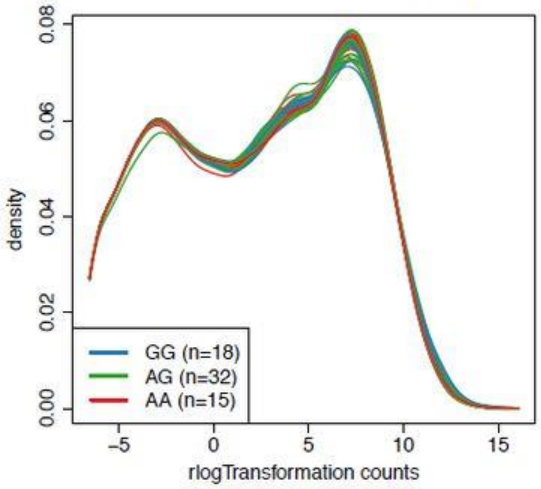
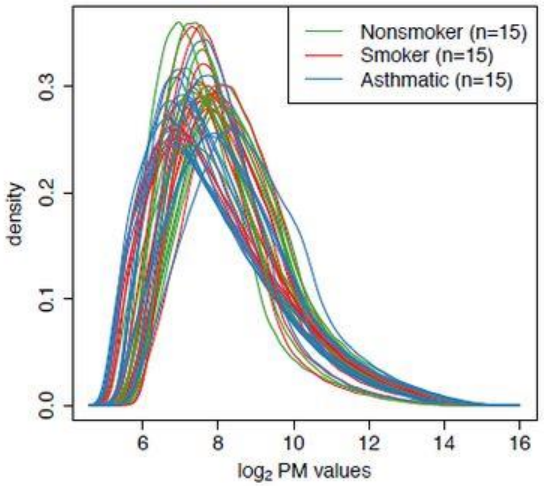
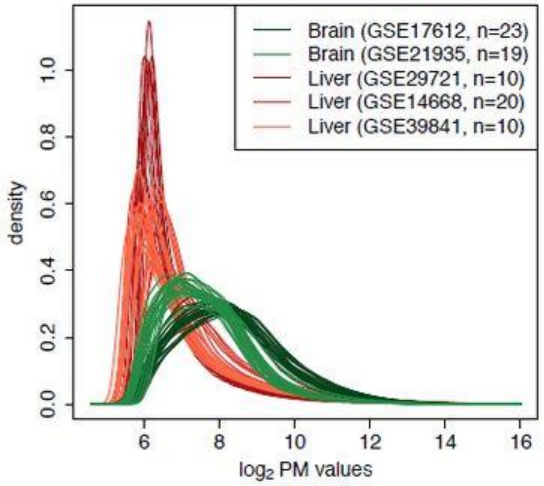


Fig. 1

A schematic of quantile normalization. Quantile normalization is a non-linear transformation that replaces each feature value (row) with the mean of the features across all the samples with the same rank or quantile. To quantile normalize a raw high-throughput data set with multiple samples: (1) order the feature values within each sample; (2) for each feature, average across the rows; (3) substitute the raw feature value with the average; (4) re-order the transformed values by placing in the original order

Quantile Normalization

	Targeted changes	Targeted changes	Global changes
Observed variation	Small variability within groups, Small variability across groups	Large variability within groups, Small variability across groups	Small variability within groups, Large variability across groups
			
Reason?	Small technical variability; no global changes	Large technical variability or batch effects *within* groups; no global changes	Raw data alone cannot detect difference Global **technical** variability or batch effects *across* groups Global **biological** variability *across* groups
What to do?	**Use quantile normalization (but not necessary)**	**Use quantile normalization**	**Use quantile normalization** **Do not use quantile normalization** *quantro* will detect global differences due to both technical and biological variation

Ejercicios

- 1) Obtener p-values con un t-test, Wilcoxon y Kolmogorov para:
  - A) 2 vectores de datos con distribución normal (rnorm) y diferente media, para número de muestras n= 2……20. Graficar los resultados y comparar.
  - B) Repetir (A) con 2 vectores de datos generados con la función runif.

- 2) Obtener p-values con un t-test, Wilcoxon, Kolmogorov y SAM para una base de datos GEO de su elección.
  - A) Verifique si necesita normalizar los datos con quantile normalization. Obtenga p-values antes y después de normalizar.
  - B) Compare los p-values y grafique. Comente sobre la correlación entre las distintas pruebas estadísticas y la cantidad de genes significativos.
  - C) Usando el GPL de los datos de GEO, dar una significancia biológica de los genes significativos.

- 3) Repetir el ejercicio (2) pero con una base de datos de GEO con menos o más muestras, según sea el caso.