V10 - 8. Gene Expression

- Regulation of Gene Transcription at Promoters
- Experimental Analysis of Gene Expression
- Statistics Primer
- Preprocessing of Data
- Differential Expression Analysis

Tue, Nov 19, 2019

8.1. Regulation of Prokaryotic Gene Transcription

Transcription consists of initiation, elongation and termination. Here, we will only be concerned with **initiation** and its control by additional proteins.

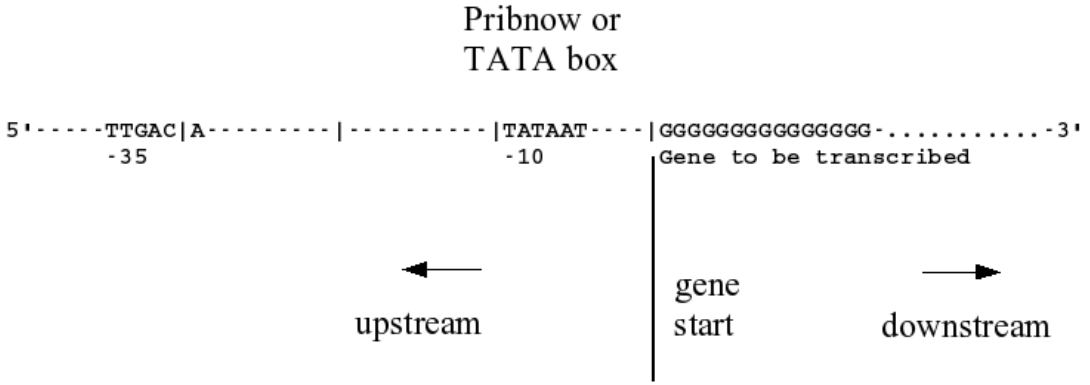
Transcription exclusively proceeds in the 5' → 3' direction.

In **prokaryotes**, transcription begins with the binding of RNA polymerase to a promoter sequence in the DNA. At the start of initiation, the bacterial core enzyme is associated with a sigma factor that aids in finding the appropriate -35 and -10 basepairs upstream of promoter sequences.

Typical promoter region of a prokaryotic gene.

The TTGACA and TATAAT motifs at positions -35 and -10 nucleotides are not essential.

The preference for the corresponding nucleotide at each position is between 50 and 80%.

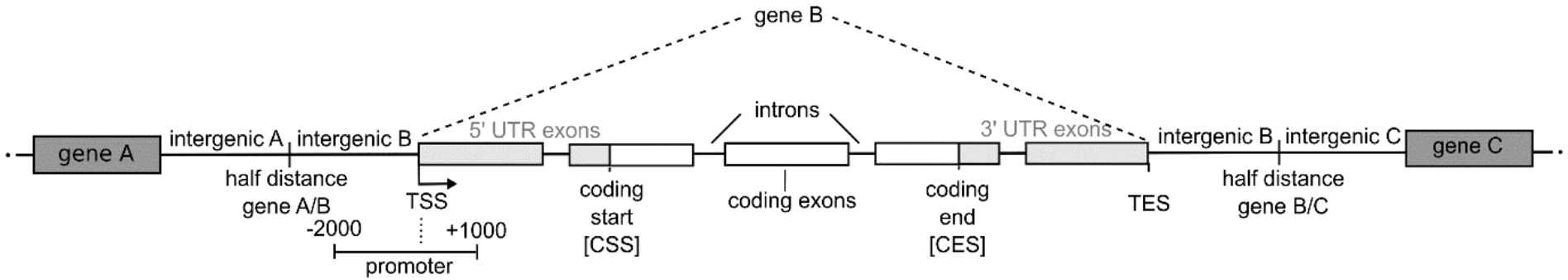


8.1. Regulation of Eukaryotic Gene Transcription

Eukaryotic transcription initiation is far more complex than in prokaryotes as eukaryotic polymerases do not recognize directly their core promoter sequences.

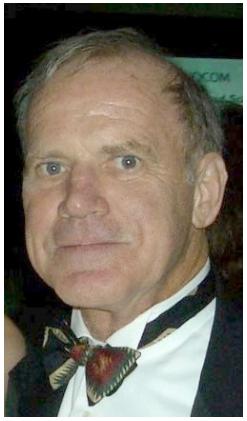
Instead, additional proteins termed **transcription factors** regulate the binding of RNA polymerase to DNA.

Many eukaryotic promoters, but not all, also contain aTATA box.



Eukaryotic genomic region containing 3 genes A, B and C. Shown in the figure is the + strand of DNA, – strand is analogous. Not shown in this figure are so-called **enhancer regions** further upstream of the promoter that play crucial roles in transcriptional regulation in higher eukaryotes.

8.2. PCR



PCR was invented in 1983 by Kary Mullis, who was awarded the 1993 Nobel Prize in chemistry for this. www.wikipedia.org

The polymerase chain reaction (PCR) is a well-known method for amplifying a specific target DNA sequence. PCR is used to isolate, sequence or clone pieces of DNA

A PCR sample consists of

- a dilute concentration of **template DNA** mixed with
- a heat-stable **DNA polymerase** (e.g. Taq polymerase),
- with **primer sequences** for the target DNA sequences,
- with deoxynucleoside triphosphates (**dNTPs**), and magnesium.

In the first step of PCR, the sample is brought to a temperature of 95–98°C.

As a result, the double-stranded template DNA **denatures** and **splits up** into two **single strands.**

8.2. PCR

In the **second step**, the temperature is lowered to about 55–65°C.

This enables the **primer** sequences to **bind** (or anneal) to complementary sequence motifs at both ends of the **target sequence** (piece from template).

In the **third step**, the temperature is usually raised to 72°C.

Then, the DNA polymerase can **extend the primer** sequences by adding dNTPs to create a new strand of DNA. Thereby, the amount of DNA is duplicated in the reaction.

This series of denaturation, annealing, and extension steps is repeated for many cycles and yields an **exponential amplification** of the template DNA.

At the end of a conventional PCR run, the amount of amplified DNA product is quantified.

8.2. Experimental Analysis of Gene Expression: real-time PCR

In „real-time“ PCR, one quantifies in real time how the amplification product accumulates during the reaction.

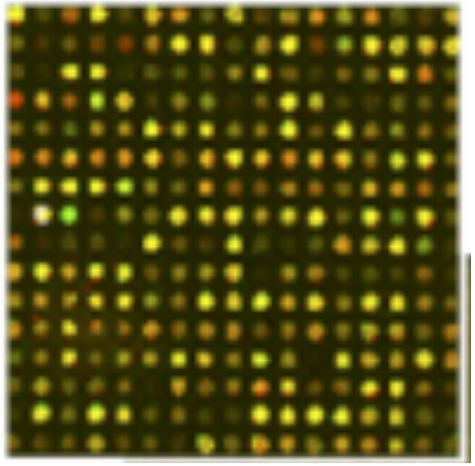
Since PCR amplifies DNA stretches, the **cellular mRNA** is first **reverse transcribed** into **cDNA** by the enzyme reverse transcriptase.

Detection of multiple PCR products in real time is made possible by adding a **fluorescent reporter molecule** to each reaction well of a parallel chip.

The detected fluorescence level is proportional to the total quantity of product DNA. The change in fluorescence over time serves to derive the amount of amplified DNA made in each cycle.

A set of multiple **internal reference genes** (e.g. suitable housekeeping genes which exhibit rather constant expression levels in all cell types and experimental conditions) is used to **normalize** the expression of target genes.

8.2. Experimental Analysis of Gene Expression: microarrays



Microarrays are a collection of DNA probes that are bound in defined positions to a solid surface, such as a glass slide.

The probes are generally oligonucleotides that are 'ink-jet printed' onto slides (Agilent) or synthesised *in situ* (*Affymetrix*).

Labelled single-stranded DNA or antisense *RNA* fragments from a sample are **hybridised** to the DNA microarray.

The amount of hybridisation detected for a specific probe is **proportional** to the number of nucleic acid fragments in the sample.

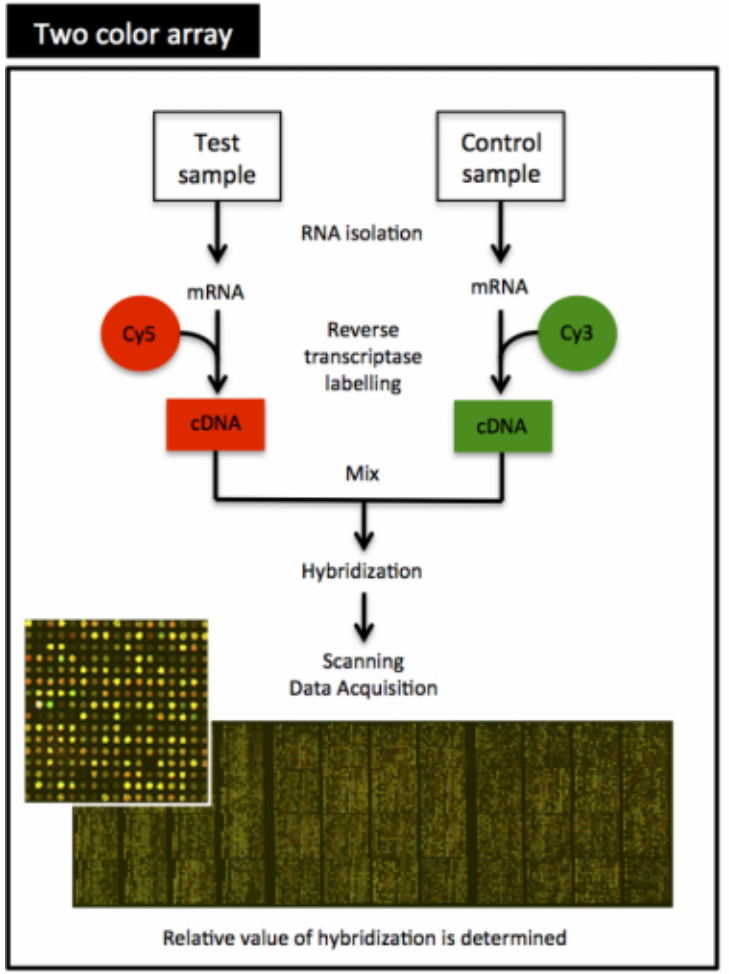
http://www.ebi.ac.uk/training/online/course/
functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

2-color microarrays

In 2-colour microarrays, 2 biological samples are **labelled** with different fluorescent dyes, usually Cyanine 3 (Cy3) and Cyanine 5 (Cy5).

Equal amounts of labelled cDNA are then simultaneously **hybridised** to the same microarray chip.

Then, the fluorescence measurements are made separately for each dye and represent the abundance of each gene in the test sample (Cy5) relative to the control sample (Cy3).



http://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

8.2. Experimental Analysis of Gene Expression: RNA-seq

The term RNA-seq describes the sequencing and determination of transcription levels of the expressed cellular mRNAome by NGS methods.

New methods are constantly entering the market every few years.

Currently, third-generation methods are state-of-the-art.

RNA-seq provides the complete genomic picture at **single-base resolution**.

We will focus here on the expression levels of entire genes.

8.3. Statistics primer

Now, we will review some basic statistics measures that are useful, e.g., when measuring gene expression using microarrays.

Given n data points denoted by a_i, where $i = 1,...,n$,

their arithmetic **mean** $\bar{a}$ (or μ) is: $$\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$$

The standard deviation σ measures how much variation or "dispersion" exists from the average (mean, or expected value).

For the same n data points $a_1, a_2, \ldots, a_n$, their **standard deviation** from the mean is:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(a_i - \bar{a}\right)^2}$$

The **variance** σ^2 is the square of the standard deviation.

8.3. Statistics primer

In probability theory, a **continuous probability distribution** *f* has to fulfil three properties:

- the probability is non-negative everywhere,

- the integral over the full distribution is normalized to one,

- and the probability that *x* lies between two points *a* and *b* is

$$p[a \leq x \leq b] = \int_a^b f(x)dx$$

8.3. Normal distribution

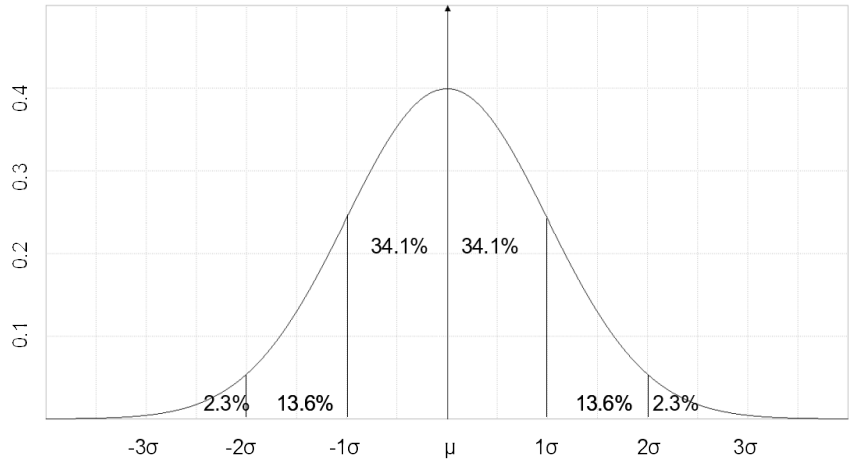
The normal (or Gaussian) distribution is a continuous probability distribution.

$$f\left(x;\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ : mean or expectation value (normally a sharp peak) and σ^2 is the variance.

The distribution with $\mu = 0$ and $\sigma^2 = 1$

is called the standard normal distribution.



Only 4.6% of the values are at least 2σ away from the mean.

This is why a **deviation of at least 2σ** is often considered a **statistically meaningful deviation**.

If a real-valued random variable clusters around a single mean value, this is typically modeled by a normal distribution as a first try.

8.3. Null hypothesis

The **null hypothesis** of a statistical hypothesis test corresponds to a general or default position.

For example, the null hypothesis might state that there is no relationship between two measured phenomena.

A null hypothesis cannot be formally proven in a mathematical sense. However, a set of data can either reject a null hypothesis or fail to reject it.

A ***p*-value** is the probability that the test statistic is at least as extreme as the one observed under the condition that the null hypothesis is true.

A small *p*-value is an indication that the null hypothesis is false.

8.3. Standard error

The standard deviation σ

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(a_i - \bar{a}\right)^2}$$

gives the „standard“ deviation of all measurements.

Often we are more interested in the standard deviation of the average.

This is denoted by the **standard error of the mean (SEM)**:

$$SEM = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(a_i - \bar{a}\right)^2}}{\sqrt{n}}$$

Whenever we use a random sample as estimate for a population, there is a good chance that our estimate will contain an error.

SEM provides an estimate for this error.

Typically, we actually need to compute SEM for the difference of the means of two random samples → 2-sample t-test.

8.3. t-tests

t-value: by how many standard errors does a difference differ from 0?

There are 3 different types of t-tests:

Unpaired t-test

$$t = \frac{average\ of\ random\ sample\ 1\ - average\ of\ random\ sample\ 2}{SEM\ of\ the\ differences\ of\ both\ averages}$$

Paired t-test

$$t = \frac{average\ of\ paired\ differences - reference\ value}{SEM\ of\ the\ differences\ of\ paired\ averages}$$

1-sample t-test

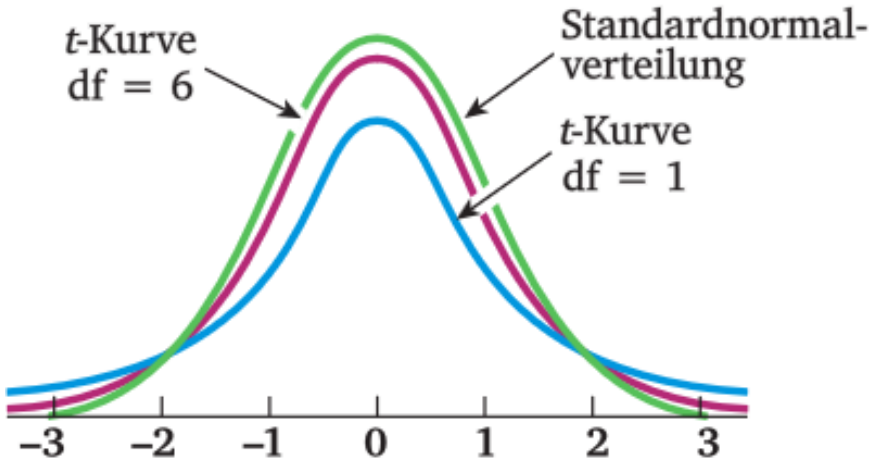
$$t = \frac{average\ of\ random\ sample\ \ - reference\ value}{SEM\ of\ the\ random\ sample}$$

https://matheguru.com/stochastik/t-test.html

8.3. t distribution

The form of the *t*-distribution is very similar to a standard normal distribution – at least for large random samples.

For small random samples, the *t*-distribution is flatter than a normal distribution.



Therefore, the *t*-distribution needs another parameter that adjusts its variance (and thus its shape).

This parameter is called the *degrees-of-freedom*; abbreviated as ***df***.

https://matheguru.com/stochastik/t-test.html

8.3. 1-sample t-test

A ***t*-test** is a **parametric** statistical hypothesis test that can be used when the population conforms to a **normal distribution**.

A frequently used *t*-test is the one-sample location *t*-test that tests whether the mean of a normally distributed population has a particular value μ_0,

$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu_0}{SEM}$$

where $\bar{x}$: sample mean,

σ : standard deviation of the sample,

n : sample size.

The **critical value** of the *t*-statistic t_0 is tabulated in *t*-distribution tables.

The hypothesis (H_0) is that the population mean equals μ_0.

If the p-value is below a threshold, e.g. 0.05, the null hypothesis is rejected.

8.3. 2-sample t-test

The 2-sample t-tests measures

$$t = \frac{average\ of\ random\ sample\ 1\ -\ average\ of\ random\ sample\ 2}{SEM\ of\ the\ difference\ of\ both\ averages}$$

Assumptions: both random samples have close to normal distribution and they have the same standard deviation.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\left(\sum X_1^2 - \frac{\left(\sum X_1\right)}{n_1} + \right)^2 + \left(\left(\sum X_2^2 - \frac{\left(\sum X_2\right)}{n_2} + \right)^2\right)}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Correction of SEM

estimated variance of X_1

Degrees of freedom

estimated variance of X_2

If 2 random variables X and Y are independent, the variance of their sum is the sum of the individual variances
$V(X+Y)=V(X)+V(Y)$

https://matheguru.com/stochastik/t-test.html

8.3. t-test

Another popular t-test is the **two sample location test**.

It tests the null hypothesis that the **mean values** of two normally distributed populations are **equal**.

Strictly speaking, the name student's t-test refers to cases when the variances of the two populations are assumed to be equal.

When this assumption is dropped, a modified test may be used that is called **Welch's t-test**.

8.3. Contingency tables

A **contingency table** is a table in matrix format that lists the (multivariate) frequency distribution of the variables.

For example, a hypothetical sample of teenagers might be grouped either by their gender (male and female) or by whether the individuals are regularly doing some sports or not. The data might look like this:

	Men	Women	Row total
Doing sports	1	9	10
Not doing sports	11	1	12
Column total	12	10	22

Obviously, there is an imbalance of men and women doing sports.

Is the **observed imbalance** in **proportions** statistically **significant**?

Such questions can be answered e.g. by **Fisher's exact test**.

8.3. Fisher's exact test

www.wikipedia.org

Fisher's exact test (named after its inventor, R. A. Fisher) is a statistical significance test that is typically used to analyze **contingency tables**.

It is valid for all sample sizes, although it is mostly used in practice when **sample sizes** are **small**.

The Fisher's exact test of independence is used if there are 2 nominal variables and we want to check whether the proportions of one variable are different depending on the value of the other variable.

The test belongs to the class of **exact tests**.

For such tests, one can compute the **significance** of deviating from a null hypothesis in an **exact way**.

For many other statistical tests, one has to rely on an approximation for the significance that becomes exact only in the limiting case of assuming an infinite sample size.

8.3. Fisher's exact test

We will now look at the example just discussed:

	Men	Women	Row total
Doing sports	1	9	10
Not doing sports	11	1	12
Column total	12	10	22

Is the **observed imbalance** in **proportions** statistically **significant**?

What is the chance probability that these 10 individuals doing sports would be so unevenly distributed between the women and the men as in this table?

We will use a **symbolic table**:

	Men	Women	Total
Doing sports	a	b	a + b
Not doing sports	c	d	c + d
Totals	a + c	b + d	a + b + c + d (=n)

8.3. Fisher's exact test

	Men	Women	Total
Doing sports	a	b	a + b
Not doing sports	c	d	c + d
Totals	a + c	b + d	a + b + c + d (=n)

The probability for any such set of values is given by the hypergeometric distribution:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\frac{(a+b)!}{a!((a+b)-a)!}\cdot\frac{(c+d)!}{c!((c+d)-c)!}}{\frac{n!}{(a+c)!(n-(a+c))!}}$$

$$= \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

This formula yields the exact probability of observing the specific distribution of the data assuming the given marginal totals, on the null hypothesis that men and women are equally likely to do sports.

The significance of any assignment of the 22 teenagers to the 4 cells of the table is obtained by considering all those cases where the marginal totals are **equally or more extreme** as those in the observed table.

8.3. Fisher's exact test

In the case study, there is only one case that is more extreme in the same direction as the given data; it is shown here:

	Men	Women	Row total
Doing sports	0	10	10
Not doing sports	12	0	12
Column total	12	10	22

Hence, we need to compute the values of p for both these tables (0.000185 and 0.00000154), and add them together (ca. 0.000187).

This corresponds to a one-tailed test.

For a two-tailed test, we must also take into account data arrangements that are equally extreme in the opposite direction.

8.3. Mann Whitney rank sum test

The Mann–Whitney U test is also called the Mann–Whitney–Wilcoxon (MWW) or Wilcoxon rank-sum test.

It belongs to the most used statistical tests among **non-parametric** statistical hypothesis testing methods.

Given a set of independent observations, this test can be used to estimate whether one sample of observations has larger values than the rest.

If the two distributions have a **different shape**, the Mann-Whitney U test is used to determine whether there are significant differences between the **distributions** of the two groups.

If the two distributions have the **same shape**, the Mann-Whitney U test is used to determine whether there are differences in the **medians** of the two groups.

8.3. Mann Whitney rank sum test

Let us assume we are given two distributions of eight values:

Class A	**2,0**	**1,7**	**1,6**	**2,1**	**2,0**	**1,4**	**2,6**	**2,2**
Class B	2,5	2,3	1,8	1,5	2,7	1,9	2,1	2,4

The values could be average grades of the pupils in two classes of a high-school or expression levels of genes A and B in several individuals.

From this, we form a joint ranked list (from lowest to highest value):

Grade	**1,4**	**1,5**	**1,6**	**1,7**	**1,8**	**1,9**	**2,0**	**2,1**	**2,0**	**2,1**	**2,2**	**2,3**	**2,4**	**2,5**	**2,6**	**2,7**
Class	**A**	B	**A**	**A**	B	B	**A**	B	**A**	**A**	**A**	B	B	B	**A**	B
Joint rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

It looks like there are more values from class A on the left side, but it is in fact impossible to judge by visual inspection of the data whether there is a significant difference between the two classes.

Thus, we will test by a rank sum test whether ranks are equally distributed in the joint rank list or not.

8.3. Mann Whitney rank sum test

Grade	1,4	1,5	1,6	1,7	1,8	1,9	2,0	2,1	2,0	2,1	2,2	2,3	2,4	2,5	2,6	2,7
Class	A	B	A	A	B	B	A	B	A	A	A	B	B	B	A	B
Joint rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

The sum of all the ranks equals $N(N + 1)/2$ where N is the total number of observations.

In this example, samples of class A have the ranks 1, 3, 4, 7, 9, 10, 11, and 15. The sum of these ranks is $T_1 = 60$.

Samples of class B have the ranks 2, 5, 6, 8, 12, 13, 14 and 16. The sum of these ranks is $T_2 = 76$.

This shows that the class B values have higher ranks on average.

8.3. Mann Whitney rank sum test

From these rank sums we compute two sums of **ranking imbalances** U, $U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - T_1$ and $U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2+1)}{2} - T_2$

n_k : number of samples in sample *k*. Here, $n_1 = n_2 = 8$.

For the example above, we get $U_1 = 40$ and $U_2 = 24$.

The correctness of these calculations can be checked by noting that the two following conditions always hold, $U_1 + U_2 = n_1 \cdot n_2$ and

$$T_1 + T_2 = (n_1 + n_2)\frac{(n_1+n_2+1)}{2}.$$

The question is how often such an imbalance in ranks can be due to chance. For this, we compare the smaller *U* value (24) with the critical value of the theoretical *U* distribution.

In this case, we get from the Mann-Whitney U-table using $n_1 = n_2 = 8$ and a significance threshold of $\alpha = 0.05$ (two-sided) a critical value of 13. Hence, the values show a significant difference between the two classes.

8.3. Kolmogorov Smirnov test

The Kolmogorov–Smirnov test (abbreviated as K–S test) is also a **nonparametric** test.

Given continuous, one-dimensional probability distributions, a one-sample K-S test compares a sample against a reference probability distribution, whereas a two-sample K-S test compares two samples to each other.

The K-S statistic determines a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples.

The two-sample KS test is one of the most useful and general nonparametric methods, as it is sensitive to differences in both **location** and **shape** of the empirical cumulative distribution functions of the two samples.

Estimate significance of differential expression (DE)

Cancer Research

Lipid Metabolism Signatures in NASH-Associated HCC—Letter

Sonja M. Kessler, Stephan Laggai, Ahmad Barghash, et al.

Cancer Res Published OnlineFirst April 28, 2014.

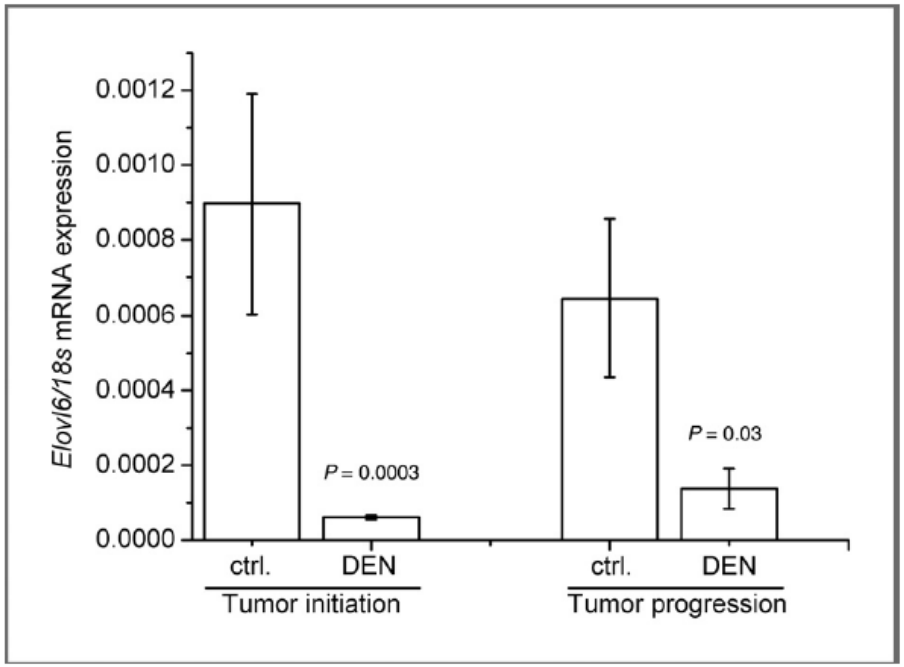


Figure 2. Wild-type mice were treated with the carcinogen DEN at the age of 2 weeks. Livers were analyzed after 24 weeks to assess the tumor initiation state. Analyses in the tumor progression stadium were done after 36 weeks. *Elovl6* mRNA expression as determined by real-time reverse transcriptase PCR with $n = 8–18$ per group. Data were normalized to *18S*. Statistical differences compared with untreated animals of the same age (ctrl.) were calculated by Mann–Whitney *U* test.

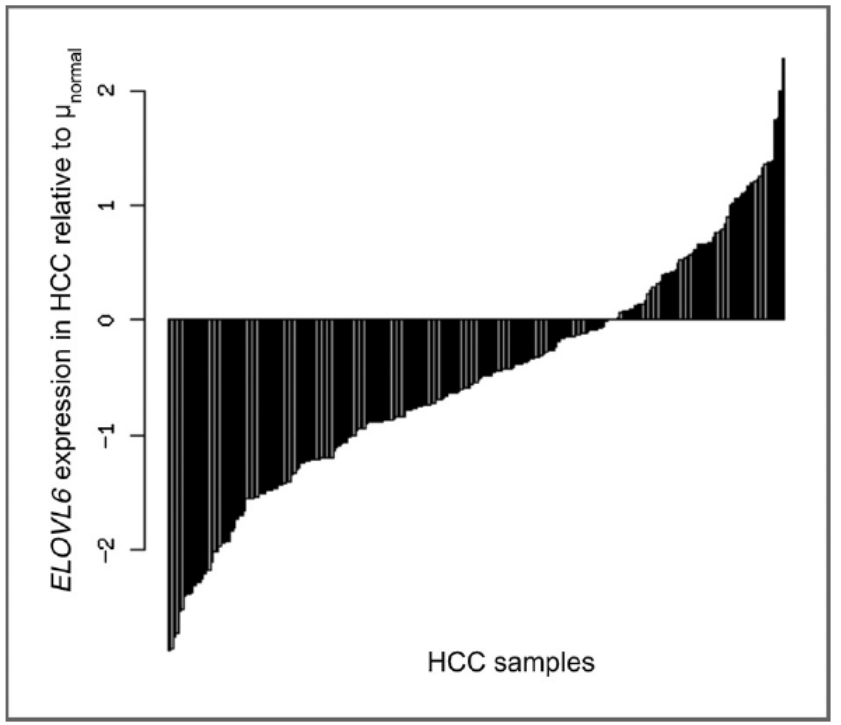


Figure 1. mRNA levels of *ELOVL6* in 247 human HCC samples relative to the mean of 239 nontumor liver tissue (μ_{normal}). Samples of dataset GSE14520 [$\log_2$ (expression) values from GEO after Robust Multi-array Average normalization] were mapped to hgu133a.db using bioconductor. Significance values: $P = 3.8E{-}11$, Kolmogorov–Smirnov test; $P = 6.7E{-}11$, *t* test; $5.1E{-}11$, Mann–Whitney *U* test.

8.3 Multiple testing problem

In hypothesis-generating studies it is a priori not clear, which terms should be tested.

Therefore, one typically performs not only one hypothesis with a single term but **many tests** with many, often all terms that the Gene Ontology provides and to which at least one gene is annotated.

Result of the analysis: a list of terms that were found to be significant.

Given the large number of tests performed, this list will contain a large number of **false-positive** terms.

Sebastian Bauer, Gene Category Analysis
Methods in Molecular Biology 1446, 175-188
(2017)

http://great.stanford.edu/

8.3 Multiple testing problem

For example, if one statistical test is performed at the 5% level and the corresponding null hypothesis is true, there is only a 5% chance of incorrectly rejecting the null hypothesis

→ one expects 0.05 incorrect rejections.

However, if 100 tests are conducted and all corresponding null hypotheses are true, the expected number of incorrect rejections (also known as false positives) is 5.

If the tests are statistically independent from each other, the probability of at least one incorrect rejection is 99.4%.

www.wikipedia.org

http://great.stanford.edu/

8.3 Bonferroni correction

Therefore, the result of a term enrichment analysis must be subjected to a **multiple testing correction**.

The most simple one is the **Bonferroni** correction. Here, each *p*-value is simply multiplied by the number of tests saturated at a value of 1.0.

The Bonferroni correction controls the so-called **family-wise error rate**, which is the probability of making one or more false discoveries.

It is a very **conservative** approach because it handles all ***p*-values** as **independent**.

Note that this is not a typical case of gene-category analysis.

So this approach often goes along with a reduced statistical power.

Sebastian Bauer, Gene Category Analysis
Methods in Molecular Biology 1446, 175-188
(2017)

http://great.stanford.edu/

8.3 Benjamini Hochberg: expected false discovery rate

The Benjamini–Hochberg approach controls the **expected false discovery rate** (FDR), which is the **proportion** of false discoveries among all rejected null hypotheses.

This has a positive effect on the statistical power at the expense of having less strict control over false discoveries.

Controlling the FDR is considered by the American Physiological Society as "the best practical solution to the problem of multiple comparisons".

Note that less conservative corrections usually yield a higher amount of significant terms, which may be not desirable after all.

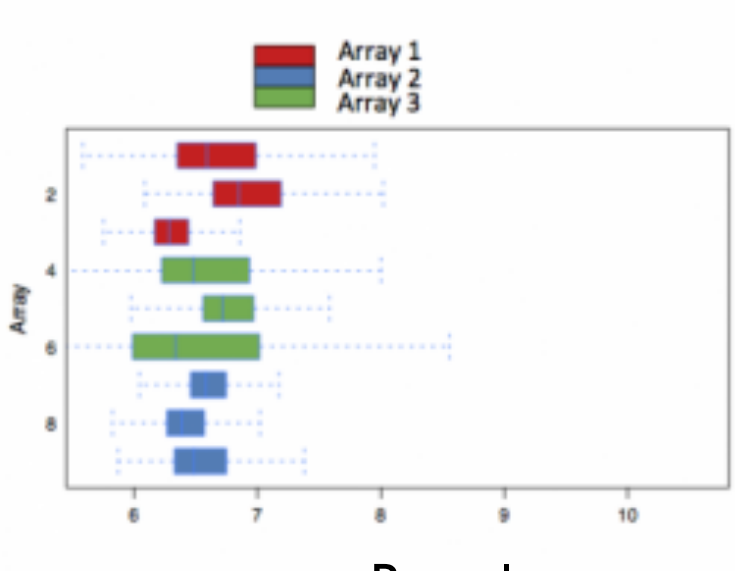
Sebastian Bauer, Gene Category Analysis
Methods in Molecular Biology 1446, 175-188
(2017)

http://great.stanford.edu/

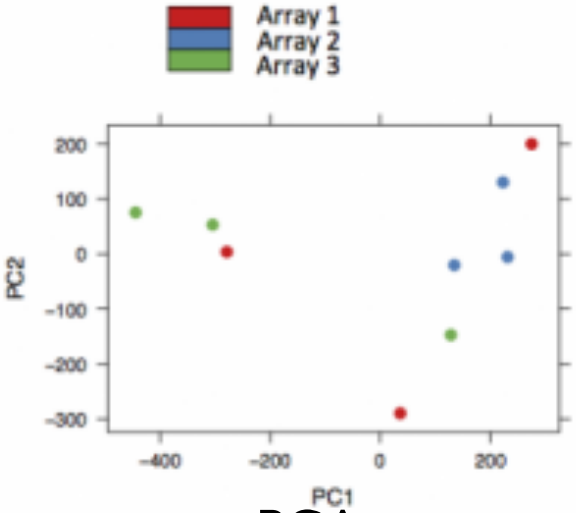
8.4 Preprocessing of data: Quality control (QC)

QC of microarray data begins with a **visual inspection** of the scanned microarray images for obvious splotches, scratches or blank areas.

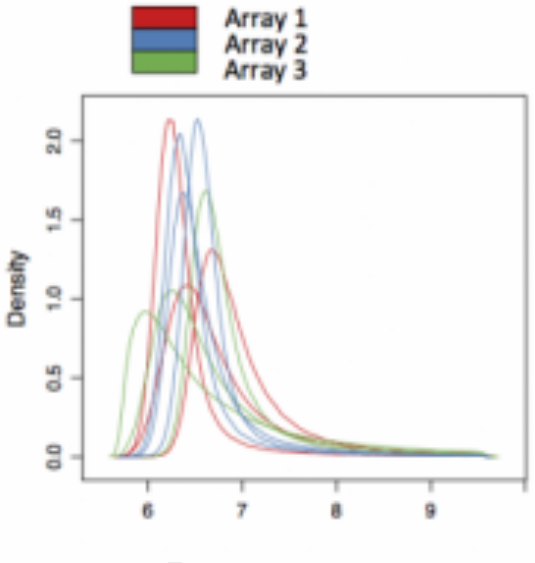
Data analysis software packages then produce different sorts of diagnostic plots, e.g. of background signal, average intensity values and percentage of genes above background to help identify problematic arrays, reporters or samples.



Box plot



PCA



Density plot

http://www.ebi.ac.uk/training/online/course/
functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

8.4 Preprocessing of data: Outliers

Analysis of expression data sets starts with identification and omission of **outlier genes** and **outlier samples**.

Outliers are experimental data points that **deviate "too much"** from the typical behaviour observed in other samples or genes.

The **reason** for such outliers could be either

- technical problems with the measurement,

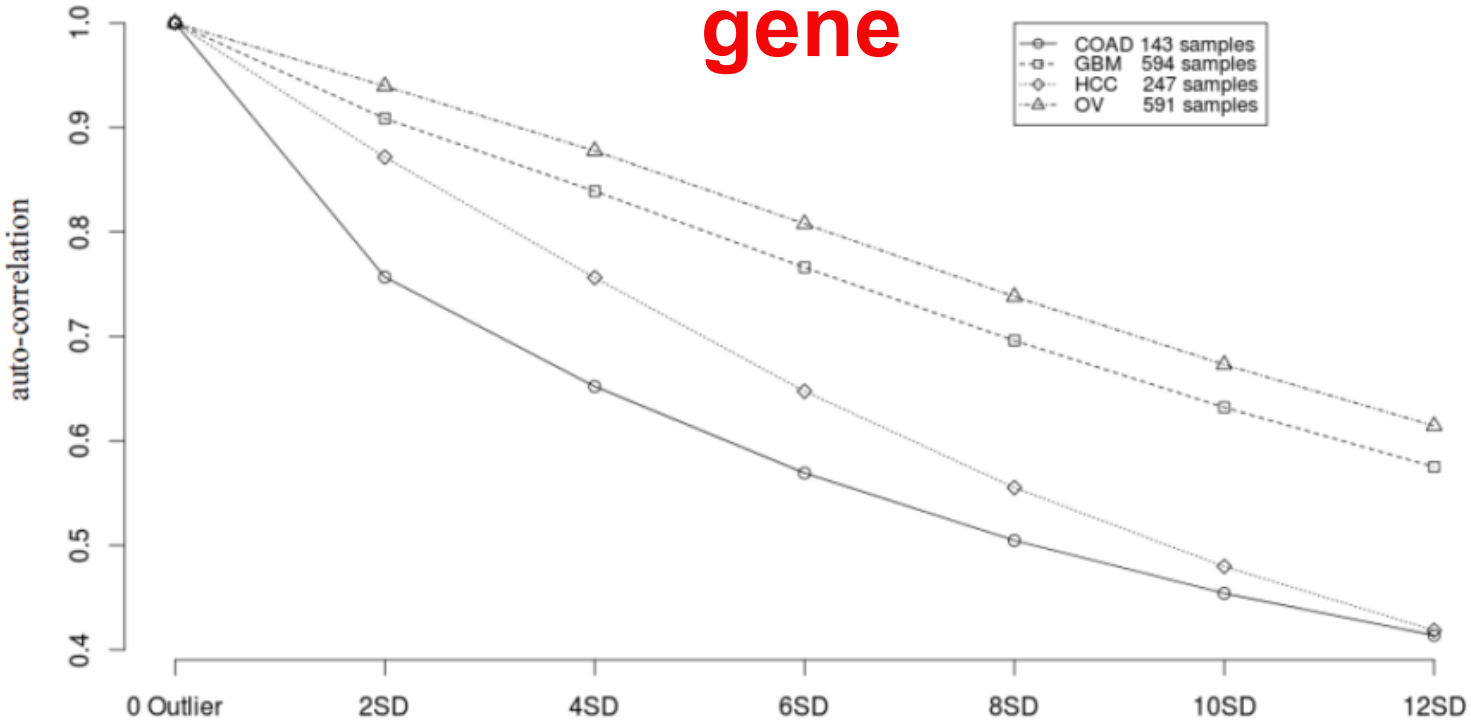
- mislabelling of samples, or

- that this sample represents a truly unique case.

Keeping such outlier data points in the data set would **obscure** the **downstream analysis**.

Typical techniques to identify outliers are hierarchical clustering, boxplots, and computing the Median Absolute Deviation (MAD) or the Generalized Extreme Studentized Deviate test (GESD).

Effect of 2 outliers on auto-correlation of a gene



Effect of 2 introduced outlier points on co-expression analysis of a gene with itself

(4 datasets from TCGA for COAD; GBM; HCC, OV tumor with hundreds of samples each).

X-axis : magnitude of perturbations applied as multiples of standard deviations (SD).

For the smallest sample (COAD), two 2SD outliers reduce the correlation to 0.75!

8.4 Detect outliers with GESD

GESD (Rosner 1983) is meant to identify one or more outliers in a dataset assuming that the majority of its data points are **normally distributed**.

For every data point *i*, the algorithm calculates its deviation from the mean μ relative to the standard deviation σ:

$$R_i = \frac{max_i|x_i - \mu|}{\sigma}$$

At each iteration, the algorithm **deletes** the **point with largest deviation**.

This process is continued until all outliers fulfilling $R_i > \lambda_i$ have been removed.

λ_i : critical values calculated for all outliers according to the *t*-distribution.

$$\lambda_i = \frac{(n-i)t_{p,n-i-1}}{\sqrt{\left(n-i-1+t^2_{p,n-i-1}\right)(n-i+1)}}$$

GESD always labels at least one outlier even when there is no outlier.

Therefore, GESD is supplied with a minimum threshold so that a certain number of outliers must be detected before any gene is marked as an outlier.

8.4 Detect outliers with MAD

In contrast to GESD, the MAD algorithm (Rousseeuw and Croux 1993) is not based on the variance or standard deviation and thus makes no particular assumption on the statistical distribution of the data.

At first, the **raw median** $median(X)$ is computed over all data points.

From this, MAD obtains the median absolute deviation (MAD) of single data points X_i from the raw median as:

$$MAD = b \cdot median(|X_i - median(X)|)$$

b is a scaling constant. For normally distributed data, one uses b = 1.4826.

As **rejection criterion** of outliers, one uses

$$\frac{X_i - median(X)}{MAD} \geq threshold$$

Suitable thresholds could be 3 (very conservative), 2.5 (moderately conservative) or 2 (poorly conservative).

8.4 Detect outliers with MAD

$$MAD = b \cdot median(|X_i - median(X)|)$$

Consider the data (1, 3, 4, 5, 6, **6**, 7, 7, 8, 9, 100).

It has a (raw) median value of 6.

The absolute deviations from 6 are (5, 3, 2, 1, 0, 0, 1, 1, 2, 3, 94).

Sorting this list into (0, 0, 1, 1, 1, **2**, 2, 3, 3, 5, 94) shows that the deviations have a median value of 2.

When scaled by b = 1.4826, the median absolute deviation (MAD) for this data is roughly 3.

Possible outliers above a rejection threshold would need to differ from the median by 6 to 9 or more.

For this example, only the extreme data point (100) deviates that much.

Normalization

Normalization is used to **control for technical variation** between assays, while **preserving the biological variation**.

There are many ways to normalize the data. The methods used depend on:
- the type of array;
- the design of the experiment;
- assumptions made about the data;
- and the package being used to analyze the data.

For the **Expression Atlas** at EBI, Affymetrix microarray data is normalised using the 'Robust Multi-Array Average' (RMA) method within the 'oligo' package.

Agilent microarray data is normalized using the 'limma' package:
'quantile normalization' for one-color microarray data;
'Loess normalization' for two color microarray data.

http://www.ebi.ac.uk/training/online/course/
functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

Quantile Normalisation

Given are: 3 measurements of 4 variables A – D.

Aim: all measurements should have an **identical distribution** of **values**.

A	5	4	3
B	2	1	4
C	3	4	6
D	4	2	8

Original data

A	iv	iii	i
B	i	i	ii
C	ii	iii	iii
D	iii	ii	iv

Determine in each column the ranks of data points.

A	2	1	3
B	3	2	4
C	4	4	6
D	5	4	8

Sort every column by magnitude

A	2	Rank i	
B	3	Rank ii	
C	4.67	Rank iii	
D	5.67	Rank iv	

Compute row averages.

A	5.67	4.67	2
B	2	2	3
C	3	4.67	4.67
D	4.67	3	5.67

Replace the original values by the averages according to the ranks of the fields.

Afterwards, all columns contain the same values (except for duplicates) and can be easily compared.

Differential expression analysis: Fold change

The simplest method to identify DE genes is to evaluate the **log ratio** between two conditions (or the average of ratios when there are replicates) and consider all genes that differ by more than an arbitrary **cut-off value** to be differentially expressed.

E.g. the cut-off value could be chosen as a **two-fold difference**.

Then, all genes are taken to be differentially expressed if the expression under one condition is over two-fold greater or less than that under the other condition.

This test, sometimes called **'fold' change**, is not a statistical test.

→ there is no associated value that can indicate the **level of confidence** in the designation of genes as differentially expressed or not differentially expressed.

Cui & Churchill, Genome Biol. 2003; 4(4): 210.

DE analysis: global *t*-test

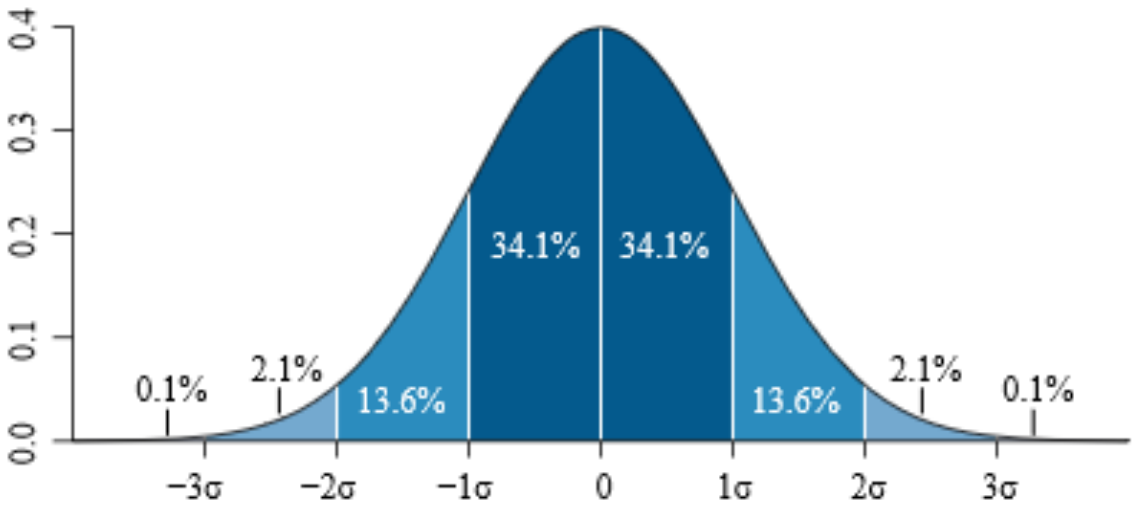
The *t* test is a simple, statistical method e.g. for detecting DE genes.

R_g : mean log ratio of the expression levels of gene g = "the effect"
SE : standard error by combining data across all genes = "the variation in the data"

Global t-test statistics : $t = \frac{R_g}{SE}$

Standard error: standard deviation of the sampling distribution of a statistic.

For a value that is sampled with an unbiased normally distributed error, the figure depicts the proportion of samples that would fall between 0, 1, 2, and 3 standard deviations above and below the actual value.



Cui & Churchill, Genome Biol. 2003; 4(4): 210;
www.wikipedia.org (M.M. Thoews)

DE analysis: gene-specific t-test

SE_g : standard error of gene g (from replicate experiments)

Gene-specific t-test statistics: $t = \frac{R_g}{SE_g}$

In replicated experiments, SE_g can be estimated for each gene from the log ratios, and a standard *t* test can be conducted for each gene.

The resulting gene-specific *t* statistic can be used to determine which genes are significantly differentially expressed.

This gene-specific *t* test is not affected by heterogeneity in variance across genes because it only uses information from one gene at a time.

It may, however, have **low power** because the sample size - the number of RNA samples measured for each condition - is typically small.
In addition, the variances estimated from each gene are **not stable**: e.g. if the estimated variance for one gene is small, by chance, the *t* value can be large even when the corresponding fold change is small.

Cui & Churchill, Genome Biol. 2003; 4(4): 210.

Differential expression analysis: SAM

As just noted, the error variance of the gene-specific *t* statistic is hard to estimate and subject to **erratic fluctuations** when sample sizes are small.

Since the square root of the **variance** gives the **denominator** of the ***t* tests**, this affects the reliability of the t-test for gene-specific tests.

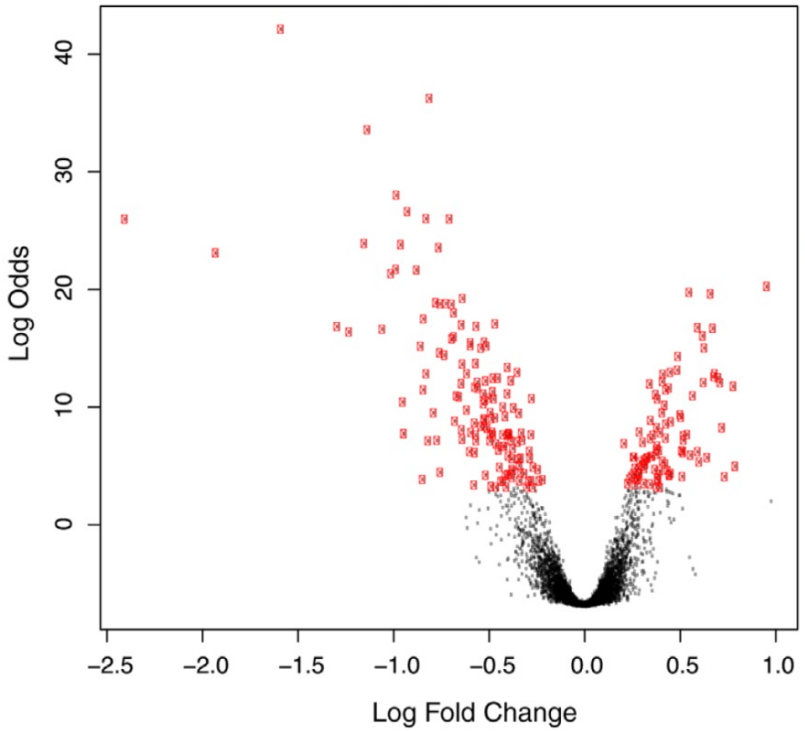
In the **'significance analysis of microarrays' (SAM)** version of the *t* test (known as the *S* test), a small positive constant *c* is added to the denominator of the gene-specific *t* test.

Significance analysis of microarrays (SAM): $S = \frac{R_g}{c+SE_g}$

With this modification, genes with small fold changes will not be selected as significant; this removes the problem of stability mentioned above.

Cui & Churchill, Genome Biol. 2003; 4(4): 210.

Limma Package: Volcano plot



The 'volcano plot' is an easy-to-interpret graph that **summarizes** both **fold-change** and ***t*-test** criteria.

It is a scatter-plot of the negative $\log_{10}$-transformed *p*-values from the gene-specific *t* test against the $\log_2$ fold change.

Genes with statistically significant differential expression according to the gene-specific *t* test will lie above a horizontal threshold line.
Genes with large fold-change values will lie outside a pair of vertical threshold lines.
The significant genes will be located in the upper left or upper right parts of the plot.

Rapaport et al. (2013) Genome Biol. 14: R95
Cui & Churchill, Genome Biol. 2003; 4(4): 210

DE detection based on RNAseq data

If sequencing experiments are considered as random samplings of reads from a fixed pool of genes,
then a natural representation of gene read counts is the **Poisson distribution** of the form

$$f(n, \lambda) = (\lambda^n e^{-\lambda})/n!$$

where *n* : number of read counts

λ : expected number of reads from transcript fragments.

An important property of the Poisson distribution
is that **variance** AND **mean** are both equal to λ, $\sigma^2 = \mu =$

However, in reality the **variance** of gene expression across multiple biological replicates is found to be **larger** than its **mean** expression values.

Rapaport et al. (2013) Genome Biol. 14: R95

DE detection in RNAseq data

To address this "**over-dispersion problem**", methods such as edgeR and DESeq use the related **negative binomial distribution** (NB) where variance σ^2 and mean μ is are related to each other by

$$\sigma^2 = \mu + \alpha\mu^2$$

where α is the "**dispersion factor**".

Different software packages (e.g. edgeR and DESeq, both by the Huber group) use different ways to **estimate** this dispersion factor.

Q: why do we need to estimate this factor?

For *k* + *r* Bernoulli trials with success probability *p*, the negative binomial distribution gives the probability of *k* successes and *r* failures, with a failure on the last trial. The values of an integer-valued random variable *K* obey to a negative binomial distribution with parameters $p \in (0, 1)$ and $r \in (0,\infty)$ if,

$$Pr(K = k) = \binom{k + r - 1}{r - 1} p^k (1 - p)^r \qquad p = \frac{\mu}{\sigma^2} \qquad r = \frac{\mu^2}{\sigma^2 - \mu}$$

Q: what is a success in our case?

r is equal to $1/\alpha$ (see above).

DESeq: detect DE genes in RNAseq data

To find the set of differentially expressed genes from RNA-Seq data modelled by a NB distribution,
mean and variance need to be estimated for each gene.

The data should be arranged as an $n \times m$ table of counts k_{ij} ,
whereby $i = 1, \ldots, n$ refers to the genes, and $j = 1, \ldots, m$ to the samples.

In DESeq, the number of reads K_{ij} in sample j that are assigned to gene i is modeled as a negative binomial distribution K_{ij} that is proportional to
$NB(\mu_{ij}, \sigma^2_{ij})$ with mean μ_{ij} and variance σ^2_{ij} .

Rapaport et al. (2013) Genome Biol. 14: R95

DESeq: detect DE genes in RNAseq data

The mean μ_{ij} is taken as $\mu_{ij} = q_{i,\rho(j)}\, s_j$

$q_{i,\rho(j)}$: expectation value of the **true concentration of fragments** from gene *i* under condition $\rho(j)$

s_j : **size factor**. It stands for the **coverage** or sampling depth of library *j*.

To estimate $q_{i,\rho(j)}$, DESeq uses the average counts from the samples *j* measured in condition ρ, after normalizing them to a common scale:

$$\hat{q}_{i\rho} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{k_{ij}}{\hat{s}_j}$$

m_ρ : number of samples in condition ρ. The sum runs over these samples.

Rapaport et al. (2013) Genome Biol. 14: R95

DESeq: detect DE genes in RNAseq data

If gene *i* is not differentially expressed or if samples *j* and *j'* are replicates, the ratio of the expected counts for this gene in different samples *j* and *j'* should match the size ratio $s_j/s_{j'}$.

Can one use the total number of reads, $\Sigma_i\, k_{ij}$, as a suitable measure of sequencing depth and set s_j equal to this number?

Based on their experience with real data, the DESeq developers argued that a few strongly and differentially expressed genes often strongly contribute to the total read count.

Hence, DESeq takes the median of the ratios of observed counts in m samples as **estimate** for the **size factors**,

$$\hat{s}_j = median_{(i)} \frac{k_{ij}}{(\prod_{\tau=1}^{m} k_{i\tau})^{\frac{1}{m}}}$$

DESeq: detect DE genes in RNAseq data

The **variance** is modeled as

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 \nu_{i,\rho(j)}$$

with the raw variance $\nu_{i\rho}$.

If one only uses the data for a single gene *i*, its variance can usually not be reliably estimated due to the small number of replicates.

Therefore, DESeq assumes that the **per-gene raw variance** parameter $\nu_{i,\rho(j)} = \nu_\rho\left(q_{i,\rho(j)}\right)$ is a smooth function of q_i and ρ and obtains $\nu_{i\rho}$ from a **fit** to the **data**.

DESeq: detect DE genes in RNAseq data

For the identification of differentially expressed genes, DESeq uses a test statistics similar to Fisher's exact test.

Let us assume a situation where we have m_A replicate samples measured in biological condition A and m_B samples measured in condition B.

The null hypothesis is that a particular gene i is expressed to the same extent in both samples, that is $q_{iA} = q_{iB}$,

q_{iA} : expression strength parameter for the samples from condition A
q_{iB} : expression strength in samples from condition B.

DESeq: detect DE genes in RNAseq data

The total counts belonging to gene *i* in each condition ρ are defined as

$$K_{iA} = \sum_{j:(j)=A} K_{ij},$$
$$K_{iB} = \sum_{j:(j)=B} K_{ij}$$

and their overall sum as $K_{iS} = K_{iA} + K_{iB}$.

Then DESeq uses any possible pairs (a, b) and their probabilities according to the modeled NB distribution, where $K_{iA} = a$ and $K_{iB} = b$ and $a + b = K_{iS}$ to calculate the p-value.

The p-value for two observed count sums (K_{iA}, K_{iB}) is obtained by adding all probabilities less or equal to $p(K_{iA}, K_{iB})$, under the condition that the overall sum is K_{iS},

$$p_i = \frac{\sum_{a+b=K_{iS}, p(a,b) \leq p(K_{iA},K_{iB})} p(a,b)}{\sum_{a+b=K_{iS}} p(a,b)}$$

.