V4 – differential gene expression analysis - outliers

V2: data imputation

V3: batch effects

- What is measured by microarrays?
- Microarray normalization
- Differential gene expression (DE) analysis based on microarray data
- Detection of outliers

- RNAseq data
- DE analysis based on RNAseq data

In today‘s lecture, we will discuss the detection of differentially expressed genes between samples from two groups.

The 2 groups may correspond to healthy and disease conditions or to two sequential stages in cellular differentiation.

Traditionally, gene expression was measured by DNA microarrays.

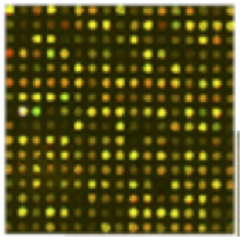
Since 2015 or so, this has been replaced more and more by next generation sequencing, namely the RNAseq technology.

But there still exists a lot of useful expression data in public repositories that was measured by microarrays.

So, bioinformaticians will keep analyzing this data in the coming years.

What is measured by 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

V4 Processing of Biological Data WS 2021/22 2

We will start with some basics about the microarray technology.

Essentially, microarrays detect the **hybridization** (binding) of single-stranded DNA stretches of the probe to single-stranded DNA probes that were chemically fixed in the wells of the microarray chip.

Each well contains many copies of the same DNA fragment.

The fragments have a typical length of 40-60 nt. If they were much shorter, then multiple DNA stretches could bind to them -> loss of specificity.

If they were much longer, this would increase the costs for production, and carry the danger that the DNA fragment finds a way to hybridize with itself -> loss of accessibility.

So if we want to apply DNA microarrays to measure the abundance of mRNAs in the sample, we first need to **reverse-transcribe** the mRNAs **into cDNA**.

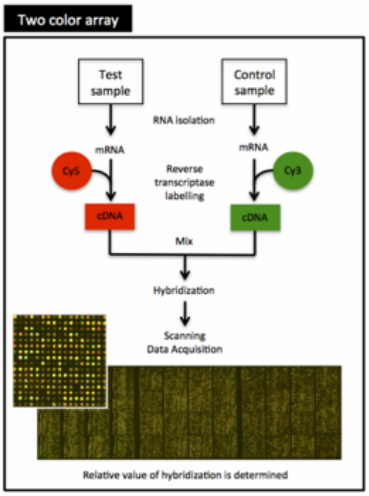
Also, we need a detectable readout. For this, we label the cDNA stretches with a fluorescent dye molecule.

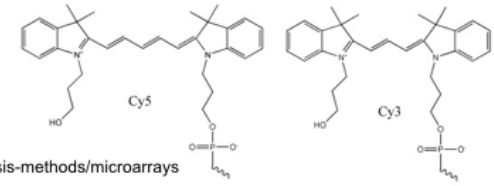
2-color microarrays

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
www.sciencedirect.com

V4 Processing of Biological Data WS 2021/22 3

If we use 2 different fluorescent dye molecules that emit at different light colors (e.g. green and red light), then we can detect to which sample the majority of cDNA/mRNA belonged to.

Remember: we are not measuring the original mRNA abundance. A cell often only contains 1 – 10 copies of individual mRNA molecules. Detecting this on a chip is practically impossible. This can only be done by mass spectrometry.

Also, we measure the amount of labeled cDNA that was obtained after several chemical processing steps. Each of them has its own efficiency.

MicroArray Quality Control (MAQC) project (2006)

MAQC project: community-wide effort that was initiated and led by FDA scientists involving 137 participants from 51 organizations.

In this project, gene expression levels were measured
- from 2 high-quality, distinct RNA samples (Universal Human Reference RNA (UHRR) from Stratagene and a Human Brain Reference RNA (HBRR) from Ambion)
- in 4 titration pools (Sample A, 100% UHRR; Sample B, 100% HBRR; Sample C, 75% UHRR:25% HBRR; and Sample D, 25% UHRR:75% HBRR.)
- on 7 microarray platforms (Applied Biosystems (ABI); Affymetrix (AFX); Agilent Technologies (AGL for two-color and AG1 for one-color); GE Healthcare (GEH); Illumina (ILM) and Eppendorf (EPP))
- and 3 alternative expression methodologies (TaqMan Gene Expression Assays; StaRT-PCR from Gene Express (GEX) and QuantiGene assays from Panomics (QGN)).

Each microarray platform was deployed at 3 independent test sites and 5 replicates were assayed at each site.

Aim of this study: find out how reproducable MA experiments are.

Nature Biotechnology **24**, 1151–1161(2006)

Here, we review the findings of a large-scale comparison that tested the reproducability of MA experiments.

This is the link to the paper on the MACS study:
https://www.nature.com/articles/nbt1239

MicroArray Quality Control (MAQC) project

The coefficient of variation (CV)

$$C_v = \frac{\sigma}{\mu}$$

relates standard deviation to mean.

Shown here is CV of the signal (not log transformed) between the intrasite replicates ($n \leq 5$) for genes that were detected in at least 3 replicates of the same sample type within a test site.

Das Bild kann derzeit nicht angezeigt werden.

Most of the one-color microarray platforms and test sites demonstrated similar replicate CV median values of 5–15%.

Nature Biotechnology **24**, 1151–1161(2006)

V4 Processing of Biological Data WS 2021/22 5

ABI – NCI are the 7 **different microarray platforms** tested. The segments labeled A to D are the 4 **titration pools**. The right system termed NCI shows higher variability.

The boxplots illustrate the coefficient of variation (y-axis left), the zig-zag lines at the top indicate the number of detected genes (y-axis right).

For each segment, there are 3 data distributions representing 3 different test sites.

The authors concluded in the abstract of their paper that there exists "intraplatform consistency across test sites".

MicroArray Quality Control (MAQC) project

Concordance of genes identified as differentially expressed for pairs of test sites, labeled as X and Y.

light-colored square: high percent overlap between the gene lists at both test sites.

dark-colored square: low percent overlap

Das Bild kann derzeit nicht angezeigt werden.

For all but the NCI test sites, the gene list overlap is at least 60% for each test site comparison (both directions) with many site pairings achieving 80% or more between platforms and 90% within platforms.

Nature Biotechnology **24**, 1151–1161(2006)

V4 Processing of Biological Data WS 2021/22 6

The authors concluded in the abstract of their paper that there exists “a high level of interplatform concordance in terms of genes identified as differentially expressed.”

We will explain in a bit how differentially expressed genes are determined by different algorithms.

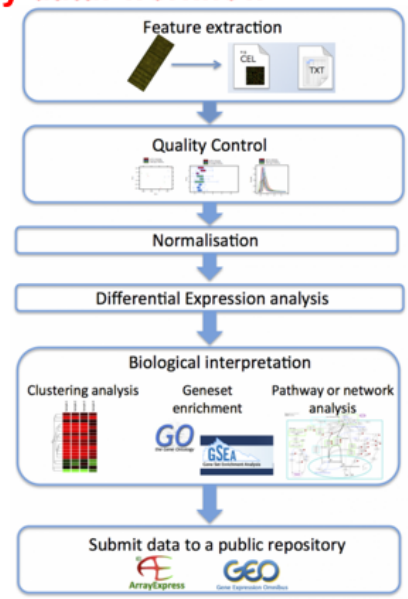
There is a follow study termed MACS-II: https://www.nature.com/articles/nbt.1665 that compared linear models for tumor outcome based on MA expression data

Analysis of microarray data: workflow

Microarrays can be used in many types of experiments including
- genotyping,
- epigenetics,
- translation profiling and
- gene expression profiling.

Gene expression profiling is by far the most common use of microarray technology.

Both one and two colour microarrays can be used for this type of experiment.



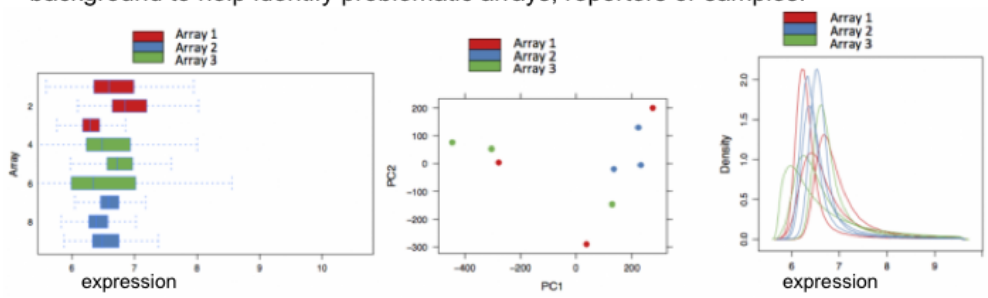
http://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

Here, we present an overview of the various steps of microarray data analysis. The individual steps listed on the flow chart will be explained on subsequent slides.

Quality control (QC) is done on the raw data

QC of microarray data begins with the **visual inspection** of the scanned microarray images to make sure that there are no obvious splotches, scratches or blank areas.

Data analysis software packages 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

Box plot, PCA and density plot are different ways to visualize the distribution of data points in the individual samples, see also lecture #2 slide 21.

In the case shown here, no apparent outlier is visible.

Normalisation

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

There are many ways to normalise 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 analyse 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 (which is based on quantile normalization).

Agilent microarray data is normalised using the 'limma' package:
'quantile normalisation' for one-colour microarray data;
'Loess normalisation' for two colour microarray data.

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

Normalization is crucial for analysis of microarray data, see also lecture #2 (quantile normalization of proteomics data).

The manufacturers of the microarray chips typically recommend particular normalization strategies that may (or may not?) be best suited for the data produced with their devices.

Usually, it is easiest to follow these instructions. This also avoids most of the trouble with reviewers of your manuscripts.

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 chosen 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.

It is not possible to give a universal threshold above which fold changes should be considered „significant“.

One aspect is statistical significance. This cannot be answered by analyzing fold changes.

Another aspect is biological *relevance*. For some genes, a small fold change may already be very relevant to the cell. For other genes, only larger fold changes may induce a phenotypic change.

Standard error of the mean

The standard deviation σ gives the „standard“ deviation of all measurements.

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

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.

V4 Processing of Biological Data WS 2021/22 11

The standard deviation measures the typical deviation of single data points from the average.

But how about the standard deviation of the average itself?

This is measured by the standard error of the mean.

It is obtained by dividing the standard deviation by the square root of the number of data points.

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

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

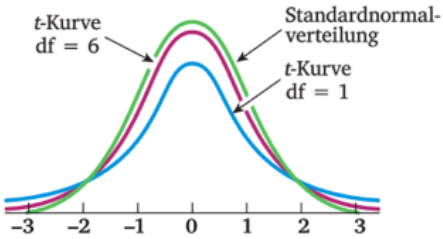
V4 Processing of Biological Data WS 2021/22 12

The student t-test compares the magnitude of the effect (e.g. what is the different of the averages of 2 sample groups) to the standard error of the mean.

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

V4 Processing of Biological Data WS 2021/22 13

To measure the statistical significance of the obtained t-values (effect over sd), the so-called t-distribution is used.

It is tabulated.

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.

The 1-sample t-test compares the mean value of a normally distributed population to a particular value.

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

Degrees of freedom

estimated variance of X_1

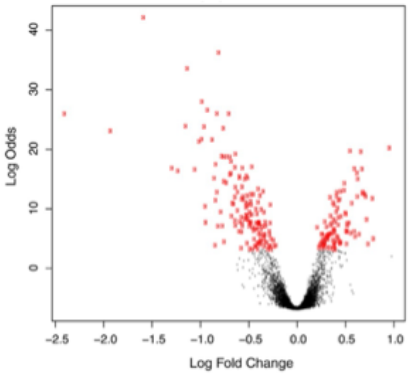
estimated variance of X_2

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

The 2-sample t-test compares the averages of two distributions.

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 identified by the *S*, *B*, and regularized *t* tests will tend to 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

V4 Processing of Biological Data WS 2021/22 16

The name of this plot reflects that the data usually has the shape of an **inverted volcano**.

Each data point is typically the difference in gene expression of one gene between samples from 2 groups, e.g. healthy vs. disease.

Each gene is characterized by its fold-change of expression (x-axis) and by the statistical significance (y-axis) that will depend on the number of samples.

Detection of Outlier Samples/Genes

Journal of Proteomics & Bioinformatics

Barghash et al., J Proteomics Bioinform 2016, 9:2
http://dx.doi.org/10.4172/jpb.1000387

Research Article Open Access

Robust Detection of Outlier Samples and Genes in Expression Datasets

Ahmad Barghash[1,2], Taner Arslan[1] and Volkhard Helms[1*]

[1]Center for Bioinformatics, Saarland University, Saarbruecken, Germany
[2]Saarbruecken Graduate School of Computer Science, Saarbruecken, Germany

Outlier : an observation that deviates "too much" from other observations.

Detecting outliers might be important either because the outlier observations are of interest themselves or because they might contaminate the downstream statistical analysis.

One common reason for outliers is **mislabeling**, where accidently a sample of one class might be falsely assigned to another one.

An outlier might also be a gene with abnormal expression values in one or more samples from the same class. In the case of cancer, this may reflect that this patient or his/her disease is a **special case**.

Now we come to the detection of outlier points.

In gene expression data, an outlier can be a problematic gene or a problematic sample.

As will be later demonstrated, it is crucially important to identify and remove problematic outlier genes/samples before the further processing of the data set.

Link to the paper: https://www.longdom.org/open-access/robust-detection-of-outlier-samples-and-genes-in-expression-datasets-jpb-1000387.pdf

Grubbs test

Grubbs' test can be used to test the presence of **one outlier** and can be used with data that is normally distributed (except for the outlier) and has at least 7 elements (preferably more).

One tests the null hypothesis that the data has no outliers vs. the alternative hypothesis that there is one outlier.

If you suspect that the maximum (minimum) value in the data set may be an outlier you can use the test statistic

$$G = \frac{x_{\max} - \bar{x}}{SD} \quad \text{or} \quad G = \frac{\bar{x} - x_{\min}}{SD}$$

The critical value for the test is

$$G_{crit} = \frac{(n-1)t_{crit}}{\sqrt{n\left(n-2+t_{crit}^2\right)}}$$

where t_{crit} is the critical value of the t distribution $T(n-2)$ and the significance level is α/n. Thus the null hypothesis is rejected if $G > G_{crit}$.

http://www.real-statistics.com/students-t-distribution/identifying-outliers-using-t-distribution/grubbs-test/

V4 Processing of Biological Data WS 2021/22 18

Grubbs' test can be used to test the presence of **one outlier** and can be used with data that is normally distributed (except for the outlier) and has at least 7 elements (preferably more).

GESD

GESD was developed to detect ≥1 outliers in a dataset assuming that the body of its data points comes from a **normal distribution**.

First, GESD calculates the deviation between every point x_i and the mean μ,

$$R_i = \frac{Max_i \left| x_i - \mu \right|}{SD}$$

normalized by the standard deviation.

At each iteration, it then removes the point with the **maximum deviation**.

This process is repeated until all outliers that fulfill the condition $R_i > \lambda_i$ are identified where λ is the critical value calculated for all points using the percentage points of the t distribution.

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

V4 Processing of Biological Data WS 2021/22 19

The Generalized Extreme Studentized Deviate (ESD) Test (Rosner 1983) is a generalization of Grubbs' Test and handles more than one outlier. It is widely used.

In GESD, you essentially run k separate Grubbs' tests to detect one or more outliers in a univariate data set that follows an approximately normal distribution.

See e.g. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h3.htm

or https://www.astm.org/standardization-news/images/nd15/nd15_datapoints.pdf

for more infos.

GESD

GESD and its predecessor ESD will **always mark** at least **one data point** as **outlier** even when there are in fact no outliers present.

Therefore, using GESD to detect outliers in microarray data must be accompanied with a **threshold** of outlier allowance where a certain amount of outliers are detected before marking a gene as an outlier.

The GESD method is said to perform best for datasets with more than 25 points.

Additionally, the algorithm requires the suspected amount of outliers as an input.

No comments.

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).

The median absolute deviation (MAD) is a measure of statistical dispersion (or variability) of the data in a population.

https://eurekastatistics.com/using-the-median-absolute-deviation-to-find-outliers/ states:

One of the most common ways of finding outliers in one-dimensional data is to mark as a potential outlier any point that is more than two standard deviations, say, from the mean.

But the presence of outliers is likely to have a strong effect on the mean and the standard deviation, making this technique unreliable.

As the standard deviation is based on *squared* distances, extreme points are much more influential than those close to the mean.

Thus it is preferential to use a measure of distance that's robust against outliers. A good candidate for this job is the *median absolute deviation from median*, commonly shortened to the *median absolute deviation* (MAD).

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 $|X_i - median(X)|$ 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.

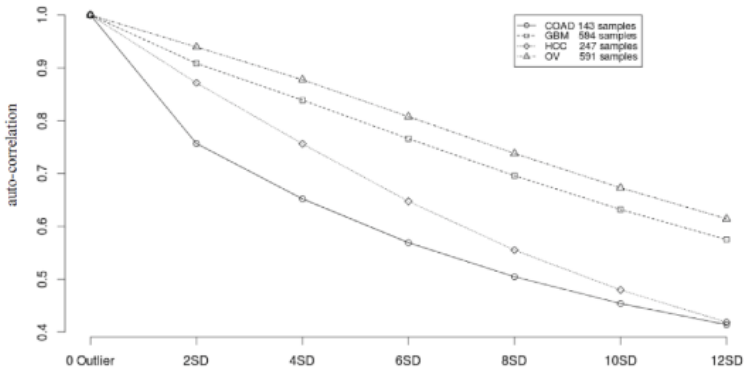
When scaled with *b* = 1.4826, the median absolute deviation (MAD) for this data is roughly 3.

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.

No comments.

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).

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

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

Barghash et al., J Proteomics Bioinform 2016, 9:2

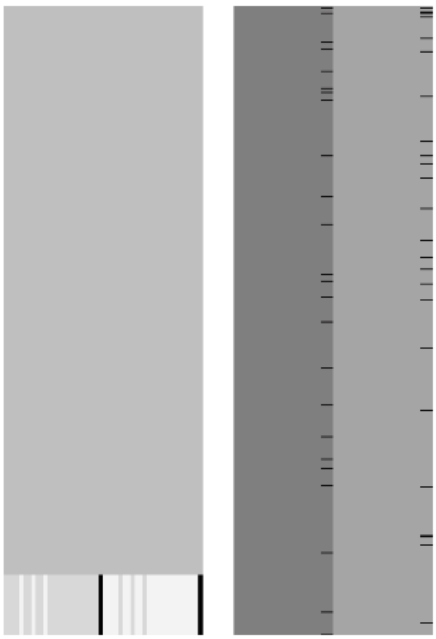
This slide shows you examples on real data sets for tumor patients from the TCGA data portal.

They are labeled COAD (for colon adenocarcinoma), GBM (glioblastoma), HCC (hepatocellular carcinoma), OV (ovarian cancer).

Measured is the auto-correlation of the expression of single genes. Without data outliers, the value should be 1.

Shown on the x-axis is the magnitude of the outlier points in multiples of standard deviation.

Simulated expression data sets



Different gray levels represent different classes.

Outlier cases are in black.

SDS1/2 (left) has **two known outliers** (**black**) and 3 known **switched samples**.

SDS3/4 (right) contain **50 outliers** each.

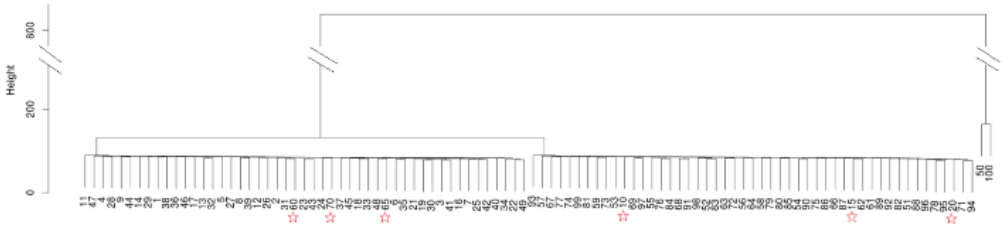
SDS1-3 follow Gaussian distributions while SDS4 follows a Poisson distribution.

Barghash et al., J Proteomics Bioinform 2016, 9:2

Here, we did a test with synthetic data that was generated by randomly drawing data points from a Gaussian distribution (SDS1-3) or from a Poisson distribution (SDS4).

Into these data sets, we introduced outlier data points of a certain magnitude at known positions.

Clustering dendogram



Clustering dendrogram of dataset of simulated expression.

Average Hierarchical Clustering based on Euclidean distances (AHC-ED) clustered SDS1 into 3 main classes grouping the **outlier samples** (50 and 100) in a separate class.

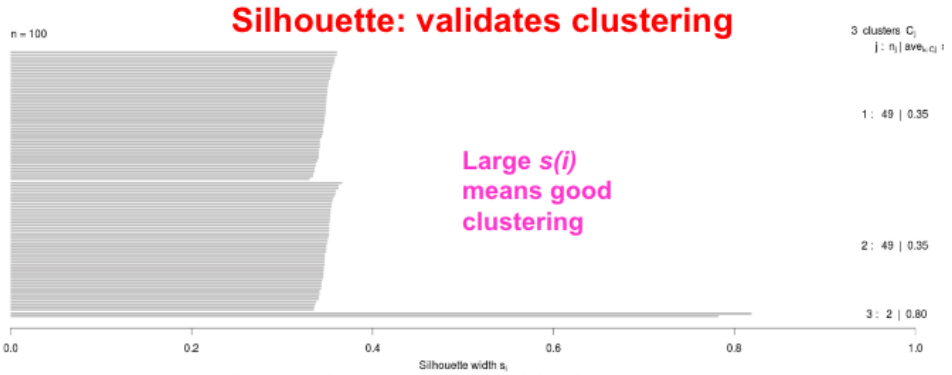
All switched samples – marked by asterisks - were **correctly clustered** into their original classes.

Barghash et al., J Proteomics Bioinform 2016, 9:2

Shown here is the clustering result.

The outliers were introduced at positions 50 and 100. This was perfectly detected by clustering.

Silhouette: validates clustering



Silhouette validation of the AHC-ED clustering of SDS1.

The average distance of 0.36 indicates that AHC-ED succeeded in clustering SDS1.

Silhouette coefficient: $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

a(i) : average dissimilarity of *i* with all other data within the same cluster
b(i) : lowest average dissimilarity of *i* to any other cluster, of which *i* is not a member

Barghash et al., J Proteomics Bioinform 2016, 9:2

This slide shows clustering of the same data as the slide before.

Shown on the x-axis is the silhouette coefficient that measures how well this data point fits into its current cluster.

A high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

of detected synthetic outlier data points (out of 50)

	GESD	Boxplot	MAD
GESD	46		
Boxplot	33	34	
MAD	33	31	33

Table 2: Detection results of simulated gene outliers.

Average of commonly detected outliers by GESD, Boxplot, and MAD algorithms in 100 simulated datasets of the SDS3 form. An outlier is considered as correctly detected if four out of five outlier values are detected from the other 50. DS3/4 has in total 50 outlier genes out of 1000.

Top: In normally distributed data, GESD identified largest number (46/50) of synthetic outliers.

Approximate Intersection	Class' Distributions	Outlier distribution	Detection Result
1SD	C1: $N(0,2^2)$ C2: $N(5,1^2)$	C1: $N(10,2^2)$ C2: $N(11,1^2)$	GESD: 45 Boxplot: 37 MAD: 36
2SD	C1: $N(0,2^2)$ C2: $N(5,1^2)$	C1: $N(8,2^2)$ C2: $N(10,1^2)$	GESD: 30 Boxplot: 18 MAD: 17
3SD	C1: $N(0,2^2)$ C2: $N(5,1^2)$	C1: $N(6,2^2)$ C2: $N(9,1^2)$	GESD: 10 Boxplot: 4 MAD: 4

Table 3: Distributions of simulation datasets.
Lists of all distributions used in different runs creating matrices of simulated expression.

Bottom: If the two distributions have larger overlap (1 SD → 2 SD →3 SD), detecting outliers becomes considerably harder.

Barghash et al., J Proteomics Bioinform 2016, 9:2

We compared the three algorithms GESD, MAD, and Boxplot in terms of their ability to identify simulated outliers in 100 generated datasets in the form of SDS3.

Each outlier gene was modeled to have 5 known outlier values out of 50 points.

The GESD algorithm was able to detect at least four out of five outlier values in 46 out of 50 outlier genes on average.

In contrast, MAD and Boxplot on average detected four out of five outlier points in only 33 and 34 genes, respectively, and some outlier points of the other outlier genes.

On average, 31 outlier genes were commonly detected by all algorithms.

MA quality control

Genomics 95 (2010) 138–142

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Minireview

Microarray data quality control improves the detection of differentially expressed genes

Audrey Kauffmann *, Wolfgang Huber

EMBL European Bioinformatics Institute, Welcome Trust Genome Campus, Hinxton, CB10 1SD, UK

These authors compared four strategies of data analysis :

- Strategy 1 No outlier removal
- Strategy 2 Outlier removal guided by arrayQualityMetrics (outliers of **boxplot**)
- Strategy 3 Removing random arrays (same number of arrays as in strategy 2)
- Strategy 4 Array weights using the function arrayWeights from the limma Bioconductor package

Kauffman, Huber (2010) Genomics 95, 138

Wolfgang Huber from EBI is the developer of several important software packages for detecting differential expression, e.g. DESeq and DESeq2.

He is also on the advisory board of the Bioconductor initiative.

Here, they analyzed whether removing outliers improves the detection of differentially expressed genes.

Link for this paper:
https://www.sciencedirect.com/science/article/pii/S0888754310000042

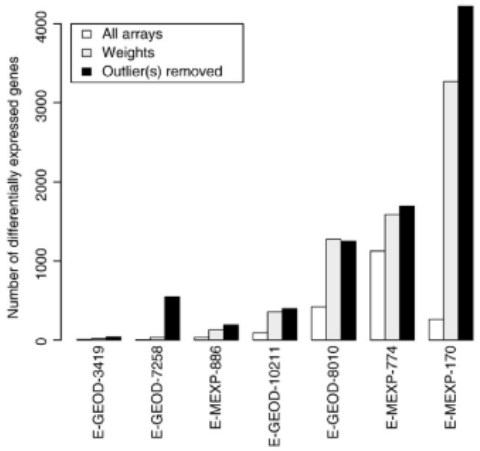
The developers of the arrayWeights method argued in
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-261

„that "bad" arrays are usually not entirely bad. Very often the lesser quality arrays do contain good information about gene expression but which is embedded in a greater degree of noise than for "good" arrays. “

In their method, an array with $\exp \gamma_j = 2$ is twice as variable as a typical array and will be given half weight in an analysis.

Number of DE genes

Data -> rma -> DE genes with moderated t-test in limma, FDR correction



Kauffman, Huber (2010) Genomics 95, 138

Number of differentially expressed genes identified:

- on the whole dataset (white bars),
- after removing outliers identified by arrayQualityMetrics (**black bars**) and
- using weights obtained by arrayWeights from limma (**grey bars**).

→ Many more DE genes identified after removing outlier genes.

E-MEXP-170 has additional confounding effect of experiment date! This explains high # of DE genes.

V4 Processing of Biological Data WS 2021/22 29

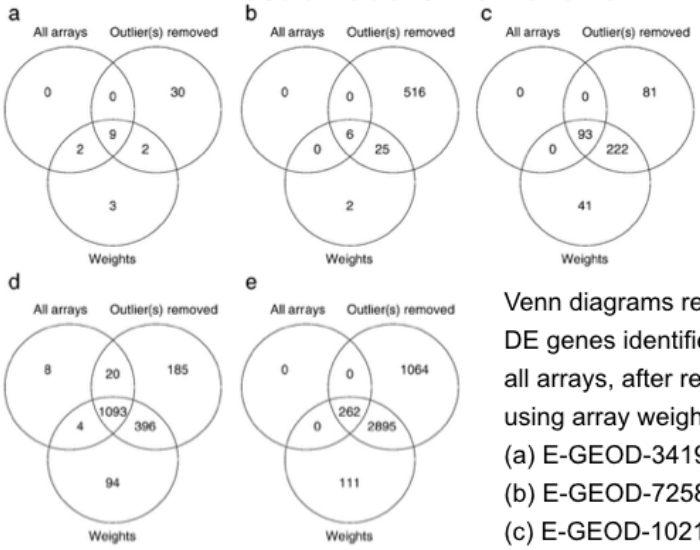
Here, the authors analyzed 7 experimental data sets.

If all data points are used (white bars), only few genes are detected as differentially expressed.

If they remove outliers identified by boxplots (black bars), the largest number of DE genes is detected.

E-MEXP-170 with over 4000 DE genes likely suffers from a confounding effect of treatment or experiment date.

Effect of Outlier removal on DE genes



In (c), (d), (e) good overlap of outlier removal and weight method.

Venn diagrams representing the number of DE genes identified by each method: all arrays, after removing outlier arrays, using array weights.
(a) E-GEOD-3419,
(b) E-GEOD-7258,
(c) E-GEOD-10211,
(d) E-MEXP-774,
(e) E-MEXP-170.

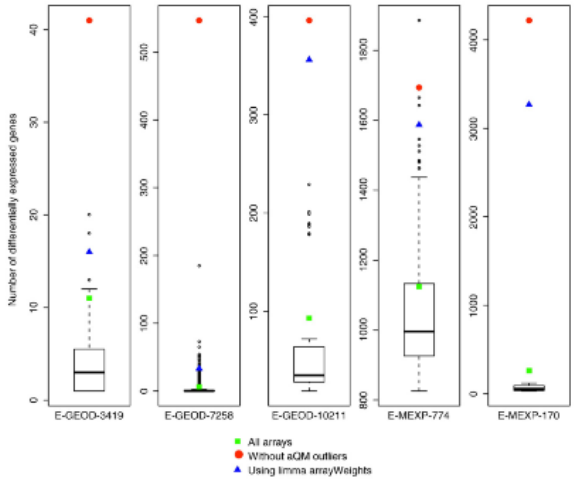
Kauffman, Huber (2010) Genomics 95, 138

The previous slide only showed that the number of DE genes increases when outliers are removed.

Does one also find the same genes?

With the exception of experiment E-GEOD-3419 (top left), the outlier removal strategy identifies almost all genes detected using the weighting method

Effect of removing random genes on DE genes



Boxplots representing the number of DE genes in each experiment when removing arbitrary subsets of size K, the number of outlier arrays identified from the N samples.

When N over K < 1000, all possible subsets were considered, otherwise 1000 subsets were sampled randomly.

If the same number of **random genes** is removed, fewer DE genes are detected.

Kauffman, Huber (2010) Genomics 95, 138

Compared with using all arrays, removal of random arrays leads to a loss of power and hence fewer genes are detected. In contrast, outlier removal and array weighting increased the numbers of differentially expressed genes.

KEGG pathway enrichment analysis

Does removal of outliers result in better biological sensitivity?

Pathway name	Genes	p-value when removing outliers	p-value when all arrays
E-GEOD-3419			
Pyrimidine metabolism	37	$<10^{-3}$	0.701
Base excision repair	17	0.001	0.542
DNA replication	19	0.003	0.451
Cell cycle	69	0.009	0.387
TGF-beta signaling pathway	48	0.009	0.558
E-GEOD-7258			
Pentose phosphate pathway	13	0.003	0.588
Fructose and mannose metabolism	28	0.003	0.326
Biosynthesis of steroids	20	0.003	0.012
Oxidative phosphorylation	44	0.003	0.299
Starch and sucrose metabolism	16	0.003	0.317

gene set enrichment analysis :

5 most enriched KEGG pathways among DE genes for experiments E-GEOD-3419 and E-GEOD-7258, with and without outlier removal.

→ The pathways are related to the biology studied in the experiments.

→ Their enrichment is **more significant** after outlier removal.

Kauffman, Huber (2010) Genomics 95, 138

Listed are the biological pathways that are enriched in DE genes.

From the biological design of the experiment, these findings are to be expected.

However, one finds them only to be significant after removing the problematic sample outliers.

Results from other outlier detection methods

ArrayExpress ID	arrayQuality Metrics	GESD	Hampel
E-GEOD-3419	6, 12	3, 6, 12	12
E-GEOD-7258	7, 15, 16	7, 15, 16	7, 15, 16
E-GEOD-10211	2, 7	2, 7	2
E-MEXP-774	4, 17	4, 17	4, 17
E-MEXP-170	6	6	6

Comparison of different outlier detection methods:

- method implemented in arrayQualityMetrics (based on **boxplots**),
- generalized extreme studentized deviate (**GESD**),
- method of Hampel (it is based on the median absolute deviation (**MAD**)).

The results of different methods overlap mostly -> robustness

Kauffman, Huber (2010) Genomics 95, 138

GESD and MAD identified very similar problematic samples.

DE analysis from RNAseq data

Compared to microarrays, RNA-seq has the following advantages for DE analysis:

- RNA-seq has a **higher sensitivity** for genes expressed either at low or very high level and **higher dynamic range** of expression levels over which transcripts can be detected (> 8000-fold range).
It also has **lower technical variation** and **higher levels of reproducibility**.

- RNA-seq is not limited by prior knowledge of the genome of the organism.

- RNA-seq detects transcriptional features, such as novel transcribed regions, alternative splicing and allele-specific expression at **single base resolution**.

While Microarrays are subject to **cross-hybridisation** bias,
RNA-seq may have a **guanine-cytosine content bias** and
can suffer from **mapping ambiguity** for paralogous sequences.

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

V4 Processing of Biological Data WS 2021/22 34

As mentioned before, the RNAseq technique has replaced microarrays since several years.

Importantly, RNAseq provides much more information about individual samples, because it also detects sequence mutations, isoforms etc.

It can be applied to novel organisms without reference genome and without availability of a standardized chip.

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

Unfortunately, the methodology for detecting DE genes from RNAseq data is not as mature yet as for microarray data.

One clear point is that assuming a Poisson distribution for the observed read counts is too unflexible in that both variance and mean must be equal to λ.

This is not observed in reality.

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.

For more details on DESeq, see Bioinformatics III lecture #10.

For the identification of differentially expressed genes, DESeq uses a test statistics similar to Fisher's exact test.
However, DESeq was found to be „overly conservative".
This led to the development of DESeq2.

The variance of data points is also termed „dispersion".

Thus, if the variance is greater than the mean, one speaks of „over-dispersion".

One way of modelling their dependence is by a polynomial with linear and quadratic term. The „dispersion factor" alpha describes the magnitude of the quadratic term.

Reference data: gold standard

Samples from **group A** : Strategene Universal Human Reference RNA (UHRR): total RNA from ten human cell lines.

Samples from **group B**: Ambion's Human Brain Reference RNA (HBRR).

ERCC **spike-in control** : mixture of 92 **synthetic** polyadenylated **oligonucleotides**, 250 to 2,000 nucleotides long, which resemble human transcripts.

The two ERCC mixtures in groups A and B contain different (known!) concentrations of 4 subgroups of the synthetic spike-ins.
Then the log expression change is predefined and can be used to **benchmark** DE performance.

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

How should one decide which differential expression analysis method is the best one?

This can only be done based on a gold-standard dataset when the correct answer is known.

But it is usually not known what genes are differentially expressed. This is what we expect from the method.

One suitable strategy is to add **synthetic data points** with known concentrations.

Here, the authors added quantities of 92 synthetically generated oligonucleotides (250 – 2000 nt long) to the probes.

This strategy is termed „**spike-in**".

These 92 oligonucleotides are then used as gold-standard set.

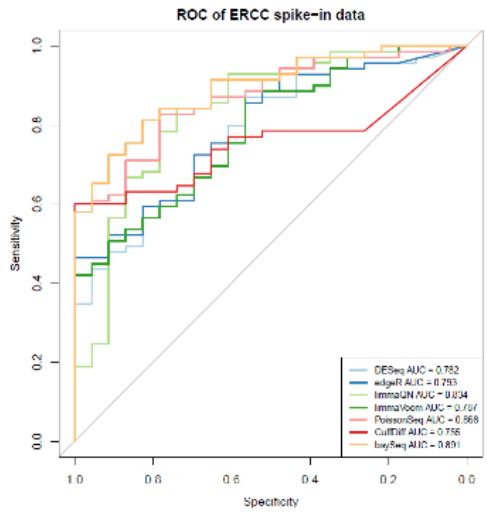
Link to this paper:
https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-9-r95

Performance for DE detection

ERCC control oligonucleotides were divided into four groups with different mixing ratios between samples A and B (1:1, 4:1, 1:2 and 2:3).

In this ROC analysis the 1:1 mix are the set of undifferentiated controls (true negatives) and all others are differentiated (true positives). AUC = area under the curve.

All methods performed reasonably well in detecting the truly differentiated spike-in sequences with an average area under the curve (AUC) of 0.78



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

This test on spike-in probes was successful, but an AUC of 0.78 is far from perfect.

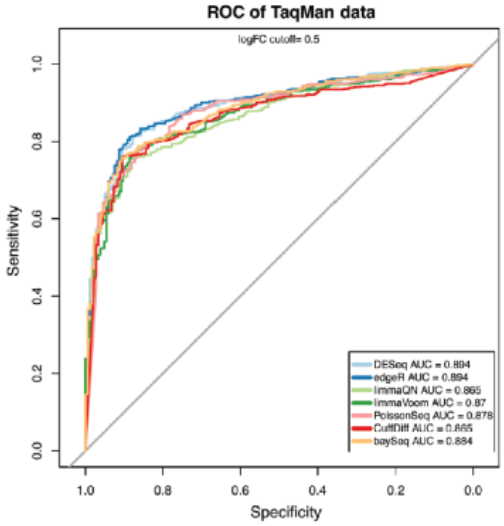
Maybe this is due to the medium size of the data set and the definition of the two classes (undifferentiated 1:1 and differentiated which contains all other mixing ratios).

Performance for DE detection

Differential expression analysis using qRT-PCR validated gene set of about 1000 genes from the MACQ project (slides 4-6).

ROC analysis was performed using a qRT-PCR $\log_2$ expression change threshold of 0.5.
If the change is >0.5, the gene is DE, otherwise not.

The results are quite comparable. **DESeq** and **edgeR** have slightly higher detection accuracy.



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

Here, the authors used a larger set of 1000 genes from the MACQ benchmark and the expression values determined by rtPCR.

Differential expression was determined based on the log2-transformed data.

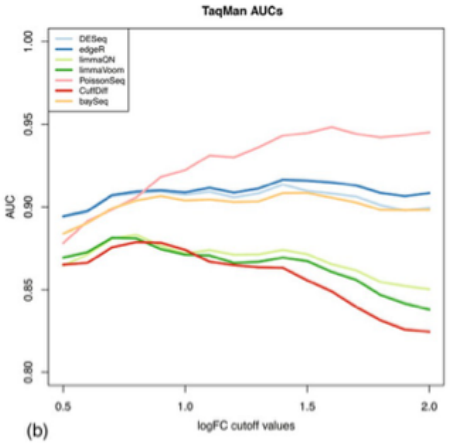
Now, all AUC values are quite good (between 0.86 and 0.89) and similar to eachother.

Performance for DE detection

If one measures AUC at increasing cutoff values of qRT-PCR expression changes, this should define sets of DE genes at increasing stringency.

Now, there is a significant performance advantage for negative binomial and Poisson-based approaches with consistent AUC values close to 0.9 or higher.

On the other hand, Cuffdiff and limma methods display decreasing AUC values indicating reduced discrimination power at higher expression change log values.



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

This test shows that one should not compare methods only at one fixed threshold.

Probably such methods are preferable that show a consistently high performance over a range of parameters.

Current situation: detecting DE genes from RNAseq data

Normalization of RNA-seq read counts is an essential procedure that corrects for non-biological variation of samples due to library preparation, sequencing read depth, gene length, mapping bias and other technical issues.

There are many normalization methods to correct for technical variations and biases: Some methods correct for read depth and transcript length:

RPKM (Reads Per Kilobase per Million mapped reads) – used by package DEGSeq

$$RPKM\ of\ a\ gene = \frac{Number\ of\ reads\ mapped\ to\ a\ gene \times 10^3 \times 10^6}{Total\ number\ of\ mapped\ reads\ from\ given\ library \times gene\ length\ in\ bp}$$

Here, 10^3 normalizes for gene length and 10^6 for sequencing depth factor.

E.g. you have sequenced one library with 5 M reads. Among them, total 4 M matched to the genome sequence and 5000 reads matched to a given gene with a length of 2000 bp.

$$RPKM\ of\ a\ gene = \frac{5000 \times 10^3 \times 10^6}{4 \times 10^6 \times 2000} = 625$$

Li et al. BMC Genomics (2020) 21:75
https://www.biostars.org/p/273537/

Link für Li-Paper:
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-6502-7

RPKM is one of the most-often used normalization methods.

Current situation: detecting DE genes from RNAseq data

FPKM (Fragments Per Kilobase per Million mapped fragments) – CuffDiff
FPKM is analogous to RPKM and used especially in paired-end RNA-seq experiments.

Other methods use global scaling quantile normalization: TC (per-sample total counts), UQ (per-sample 75% upper quartile Q3), Med (per-sample Median Q2), or Q (full quantile) implemented in Aroma.light.

DESeq/DESeq2 and edgeR use an imputed size factor to correct for read depth bias.

RUV normalizes by the expression of control genes to remove unwanted technical variation across samples.

Sailfish is an alignment-free abundance estimation using k-mers to index and count RNA-seq reads.

Li et al. presented a method called UQ-pgQ2 (per-gene Q2 normalization following per-sample upper-quartile global scaling at 75 percentile) for correcting library depths and scaling the reads of each gene into the similar levels across conditions.

FPKM is analogous to RPKM.

But there exist many other normalization methods.

Comparison of different methods

Table 1 Summary of studies comparing normalization methods for the DEG analysis

References	Normalization methods	Software Packages/ pipelines	Replicates per condition (n)	Conclusions
Bullard et al. 2010 [17]	POLR2A, Q, TC, UQ	Genominator	2, 4	POLR2A and UQ with LRT/Exact test significantly reduced the bias of DE relative to qRT-PCR
Tang et al. 2015 [40]	RLE,TMM, UQ, RPKM, FPKM, Q, voom,	DESeq, DESeq2, edgeR, EBSeq, baySeq, SAMseq, PoissonSeq, voom-limma, TCC	1, 3, 6, 9	In multi-group comparison, the proposed pipeline internally using edgeR was recommended for count data with replicates while this pipeline with *DESeq2* was recommended for data without replicates
Germain et al. 2016 [41]	RLE, TMM, voom, TPM	Cufflinks-CuffDiff, DESeq2, edgeR, voom-limma	3, 5	With benchmarked differential expression analysis, in general voom and *edgeR* showed the most stable performance and be superior to other methods in most assay with replicates of 3 and 5. But voom significantly underperformed in transcript-level simulation and edgeR shown suboptimal results in the SEQC dataset
Maza E 2016 [42]	TMM, RLE, MRN	DESeq2, edgeR	1	The three methods gave the same results for a simple two-condition comparison withourt replicates.
Costa Silva et al. 2017 [43]	TMM, RLE, UQ, voom	Limma-Voom, NOIseq, DESeq2, SAMSeq, EBSeq, sleuth, baySeq, edgeR	1:8	Limma-voom, NOIseq and *DESeq2* had more consistent results for DEGs identification
Spies et al. 2019 [44]	Vst, Med, RLE, TMM	DyNB, EBSeq-HMM, FunPat, ImpulseDE2, lmms, next maSigPro, nsgp, splineTC, timeSeq, edgeR, DESeq2	2, 3, 5	*DESeq2* and *edgeR* with a pairwise comparison outperformed TC tools for short time course (< 8 time points) due to high false positive rate except ImpulseDE2, but they were less efficient on longer time series than splineTC and maSigPro tools.

Other earlier studies were left out.

Li et al. BMC Genomics (2020) 21:75

V4 Processing of Biological Data WS 2021/22 43

There exists already a number of benchmark studies, but no consistent trends are apparent yet.

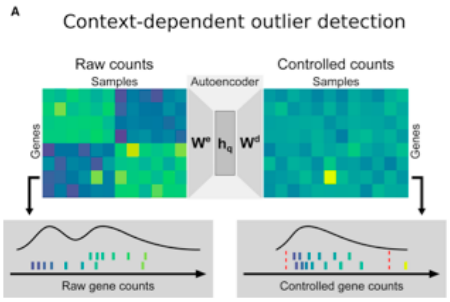
DESeq2 is often among the best-performing methods, but not always.

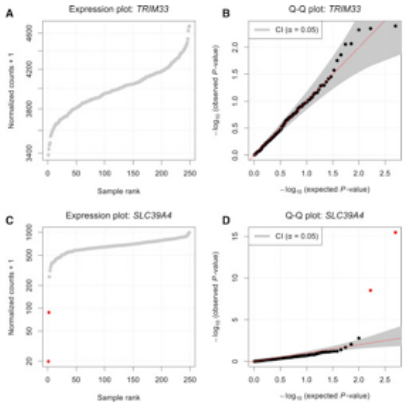
Li et al. found for the benchmark MAQC dataset that their own method performed best.

I guess the jury is still out what method will make it in the long run.

Outlier detection for RNA-seq data: Outrider

Outlier detection is equally important when processing RNA-seq data.





Based on synthetic data, an autoencoder is entrained to detect outlier data points.

Normalized RNA-seq read counts plotted against their rank (A and C) and quantile-quantile plots of observed p values against expected p values with 95% confidence bands (B and D); outliers are shown in red (FDR < 0.05). Shown are data for *TRIM33* with no detected expression outlier (A and B) and data for *SLC39A4* with two expression outliers (C and D).

Brechtmann … Gagneur,

Am J Hum Genet. (2018) 103, 907-917.

V4 Processing of Biological Data WS 2021/22 44

This paper by the group of Julien Gagneur presents a Deep Learning (autoencoder) method termed Outrider to identify outliers in RNAseq data.

The left figure illustrates schematically how the autoencoder transforms raw counts into so-called controlled counts.

Now, the yellow-colored field clearly represents an outlier that was not detectable in the raw counts.

The right figure presents two ways of representing expression data.

The upper example belongs to gene TRIM33, the lower example to the gene SLC39A4 (a membrane transporter).

For SLC39A4, two clear outliers are visible both in the sample rank plot as well as in the Q-Q plot for the p-values.

ARTICLE

Convolution of bulk sequencing data

https://doi.org/10.1038/s41467-020-14561-0 OPEN

Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants

Margaret K.R. Donovan[1,2], Agnieszka D'Antonio-Chronowska[3], Matteo D'Antonio[3*] & Kelly A. Frazer[3,4*]

Genotype-tissue expression (GTEx) project:
over 10,000 bulk RNA-seq samples representing 53 different tissues from 30 organs obtained from 635 genotyped individuals.
The aim is to link the influence of genetic variants on gene expression levels through quantitative trait loci analysis (eQTL).

Problem: data set does not account for cellular heterogeneity (i.e., different cell types within a tissue and the relative proportions of each cell type across samples of the same tissue)

Possible solution: deconvolute data into separate cell types.

Donovan et al. (2020) Nature Commun. 11:955

Link to this paper: https://www.nature.com/articles/s41467-020-14561-0

Large-scale projects such as GTEx have produced very valuable and costly datasets. However, many of these methods used bulk sequencing, not single-cell sequencing.

Can one decompose / deconvolute these data sets into the contributions of individual cell types?

Convolution of bulk sequencing data

Das Bild kann derzeit nicht angezeigt werden.

In a *proof-of-concept* analysis, the cellular estimates of 2 GTEx tissues (liver and skin) were deconvoluted using both mouse and human signature genes obtained from scRNA-seq.

We then performed *cellular deconvolution* of the 28 GTEx tissues from 14 organs using CIBERSORT and characterized both the heterogeneity in cellular composition between tissues and the heterogeneity in relative distributions of cell populations between RNA-seq samples from a given tissue.
Finally, we used the cell type composition estimates as interaction terms for *eQTL analyses* to determine if we could detect cell-type-associated genetic associations.

Donovan et al. (2020) Nature Commun. 11:955

The idea is to steer the convolution by providing a certain amount of single-cell sequencing data either from human or from mouse.

CIBERSORT

Deconvolution of gene expression profiles (GEP) can be represented by **M** = **f** × **B**, provided that **B** contains more marker genes than cell types (i.e., the system is overdetermined).

M : mRNA mixture

B : GEP signature matrix

f : vector consisting of the unknown fractions of each cell type in the mixture

Previous groups have applied linear least squares regression (LLSR) and more recently, non-negative least squares regression (NNLS) and quadratic programming (QP) to solve for **f**.

Cibersort uses *v*-support vector regression (details are not important here).

Newman et al. Nature Methods 12, 453–457 (2015)

Deconvolution was done using the CIBERSORT software that uses nu-support vector regression to split up samples into groups.

The details of nu-support vector regression are not relevant at this point.

CIBERSOFT software: https://www.nature.com/articles/nmeth.3337

Convolution of bulk sequencing data

Bar plots showing the fraction of cell types estimated in the 175 GTEx liver RNA-seq samples deconvoluted using
c gene expression profiles from high-resolution human liver scRNA-seq, or
d from low-resolution mouse liver scRNA-seq, or
e GTEx estimates generated by collapsing high-resolution human cell types within each of the seven distinct cell classes.

Hepatocyte estimates from mouse liver were positively and highly correlated with the human high-resolution hepatocyte 0 population estimate ($r = 0.71$, p-value $= 5.4 \times 10^{-28}$).

Das Bild kann derzeit nicht angezeigt werden.

Donovan et al. (2020) Nature Commun. 11:955

The upper plot shows the convolution of human bulk liver sequencing data into 15 different cell types present in human livers.

The middle plot shows a deconvolution of the same bulk data into 5 broad types of mouse liver cells.

The bottom plot shows a deconvolution of the same data when the data of the top plot is collapsed into seven broad types.

Interpretation: scRNA-seq generated from human and mouse liver captured similar cell types.

Technical differences, including the number of cells analyzed and tissue sampling methodology, affect the cell type resolution.

Summary

Removing outlier data sets from the input data is essential for the downstream analysis (unless these outliers are of particular interest -> personalized medicine).

Analysis tools: box-plots, PCA, density plots, clustering

Some outlier methods (GESD) are based on variants of the ***t*-test**.

MAD and boxplots are other simple methods.

Normalization of RNA-seq data: many different strategies exist.

Single-cell data based **deconvolution** of bulk sequencing data can help in increasing the insight that can be obtained from existing bulk data.

Additional slides (not used)

CIBERSORT uses nu–support vector regression (*v*-SVR).

v-SVR is an instance of support vector machine (SVM), a class of optimization methods for binary classification problems, in which a hyperplane is discovered that maximally separates both classes.
The support vectors are a subset of the input data that determine hyperplane boundaries. Unlike standard SVM, SVR discovers a hyperplane that fits as many data points as possible (given its objective function) within a constant distance, ε, thus performing a regression.

All data points within ε (termed the 'ε-tube') are ignored, whereas all data points lying outside of the ε-tube are evaluated according to a linear ε-insensitive loss function. These outlier data points, referred to as 'support vectors', define the boundaries of the ε-tube and are sufficient to completely specify the linear regression function.
In this way, support vectors can provide a sparse solution to the regression in which overfitting is minimized (a type of feature selection). Notably, support vectors represent genes selected from the signature matrix in this work.

Newman et al. Nature Methods 12, 453–457 (2015)

CIBERSOFT software: https://www.nature.com/articles/nmeth.3337

CIBERSORT

A simple 2D dataset analyzed with linear *v*-SVR, with results shown for two values of *v* (note that both panels show the same data points). As linear SVR identifies a hyperplane (which, in this 2D example, is a line) that fits as many data points as possible (given its objective function) within a constant distance, ε (open circles).
Data points lying outside of this 'ε-tube' are termed 'support vectors' (red circles), and are penalized according to their distance from the ε-tube by linear slack variables (ξ_i).

Das Bild kann derzeit nicht angezeigt werden.

Importantly, the support vectors alone are sufficient to completely specify the linear function, and provide a sparse solution to the regression that reduces the chance of overfitting. In *v*-SVR, the *v* parameter determines both the lower bound of support vectors and upper bound of training errors. As such, higher values of *v* result in a smaller ε-tube and a greater number of support vectors (right panel). For CIBERSORT, the support vectors represent genes selected from the signature matrix for analysis of a given mixture sample, and the orientation of the regression hyperplane determines the

Newman et al. Nature Methods 12, 453–457 (2015)

CIBERSOFT software: https://www.nature.com/articles/nmeth.3337

CIBERSORT

Das Bild kann derzeit nicht angezeigt werden.

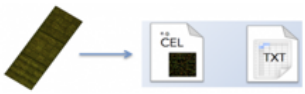
CIBERSORT requires an input matrix of reference gene expression signatures, collectively used to estimate the relative proportions of each cell type of interest. To deconvolve the mixture, we employ a novel application of linear support vector regression (SVR), a machine learning approach highly robust with respect to noise. Unlike previous methods, SVR performs a feature selection, in which genes from the signature matrix are adaptively selected to deconvolve a given mixture. An empirically defined global *P* value for the deconvolution is then determined.

Newman et al. Nature Methods 12, 453–457 (2015)

CIBERSOFT software: https://www.nature.com/articles/nmeth.3337

Extraction of features

Feature extraction is the process of **converting** the scanned image of the microarray into **quantifiable values** and annotating it with the gene IDs, sample names and other useful information



This process is often performed using the software provided by the microarray **manufacturer**.

Common microarray **raw data file types**.

Manufacturer	Typical raw data format	How to open / Analysis software examples
Affymetrix	.CEL (binary)	**R packages** (affy, limma, oligo…)
Agilent	feature extraction file (tab-delimited text file per hybridisation)	Spreadsheet software (Excel, OpenOffice, etc.)
GenePix (scanner)	.gpr (tab-delimited text file per hybridisation)	Spreadsheet software (Excel, OpenOffice, etc.)
Illumina	.idat (binary)	**R packages** (e.g. illuminaio)
	txt (tab-delimited text matrix for all samples)	Spreadsheet software (Excel, OpenOffice, etc.)
Nimblegen	NimbleScan, .pair (tab-delimited text matrix for all samples)	Spreadsheet software (Excel, OpenOffice, etc.)

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

The .CEL files produced from Affymetrix chips and the .idat from Illumina chips are most common.