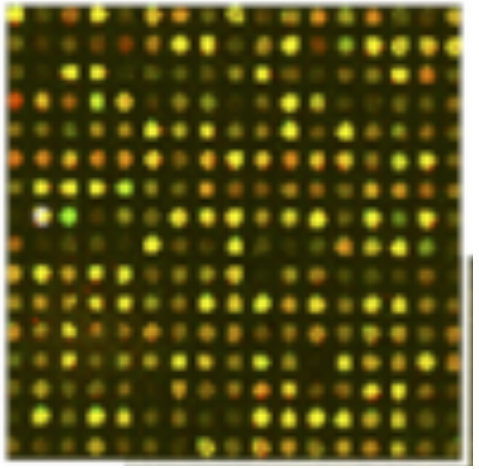
V3 – differential gene expression analysis

- 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

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.

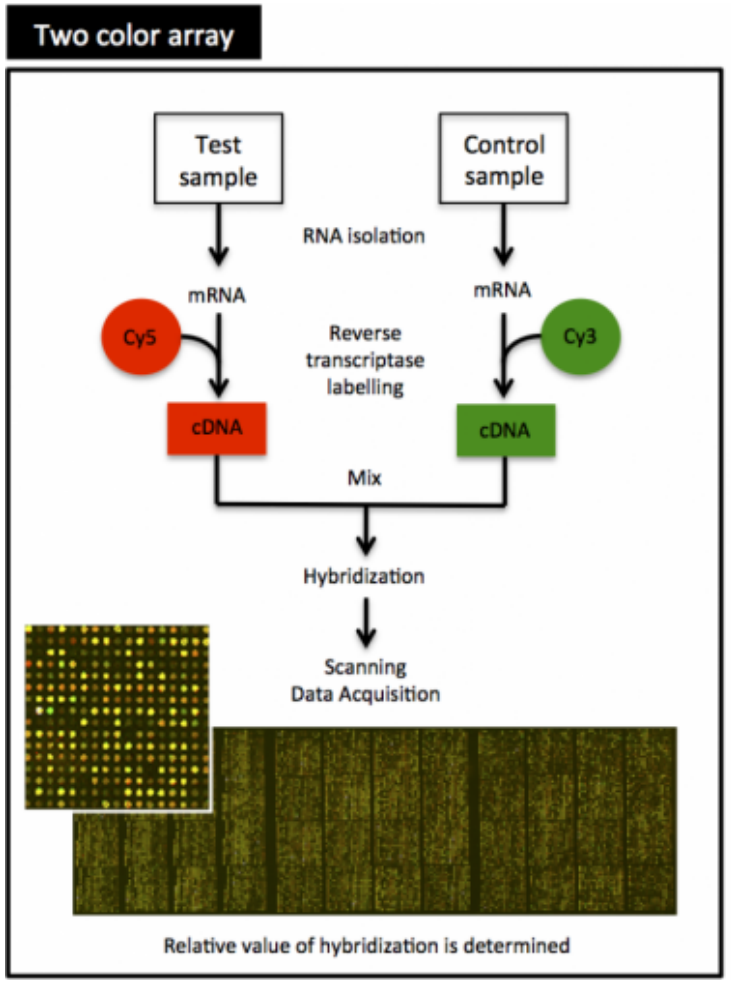
http://www.ebi.ac.uk/training/online/course/
functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

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.

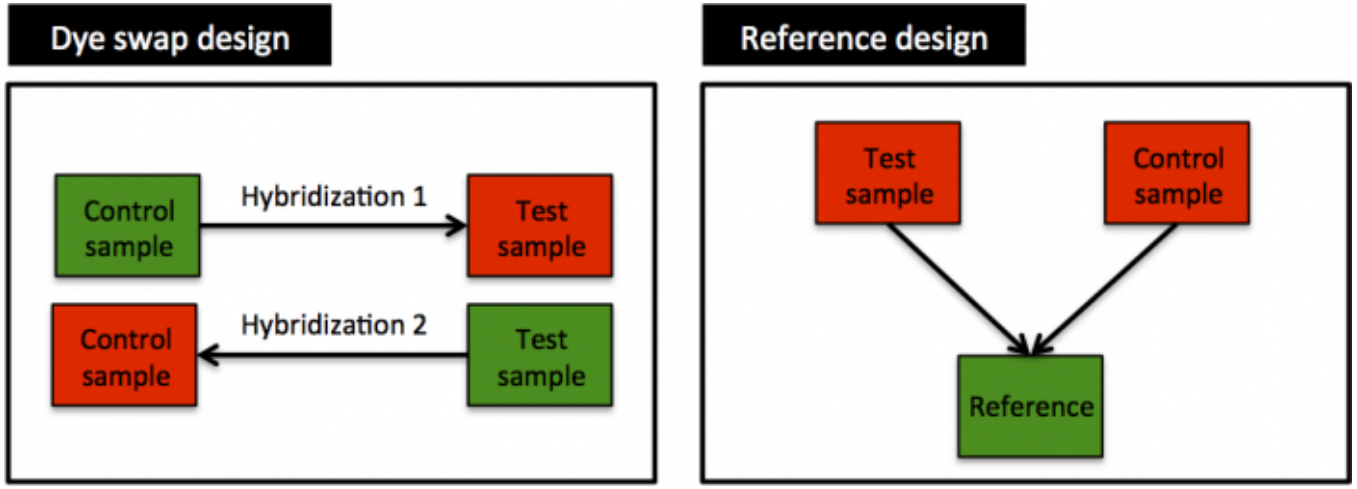
Then, the fluorescence measurements are made separately for each dye and represent the abundance of each gene in the test sample (Cy5) relative to the control sample (Cy3).



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

Counter acting dye bias

One issue for two-colour arrays is related to **dye bias effects** introduced by the slightly different **photo-chemistry** of the two dyes.
This effect can be **corrected** in 2 different ways.



In a **dye swap design**, the same pairs of samples (test and control) are compared twice with the dye assignment reversed in the second hybridisation.

The most common design for two colour microarrays is the **reference design** in which each experimental sample is hybridised against a common reference sample.

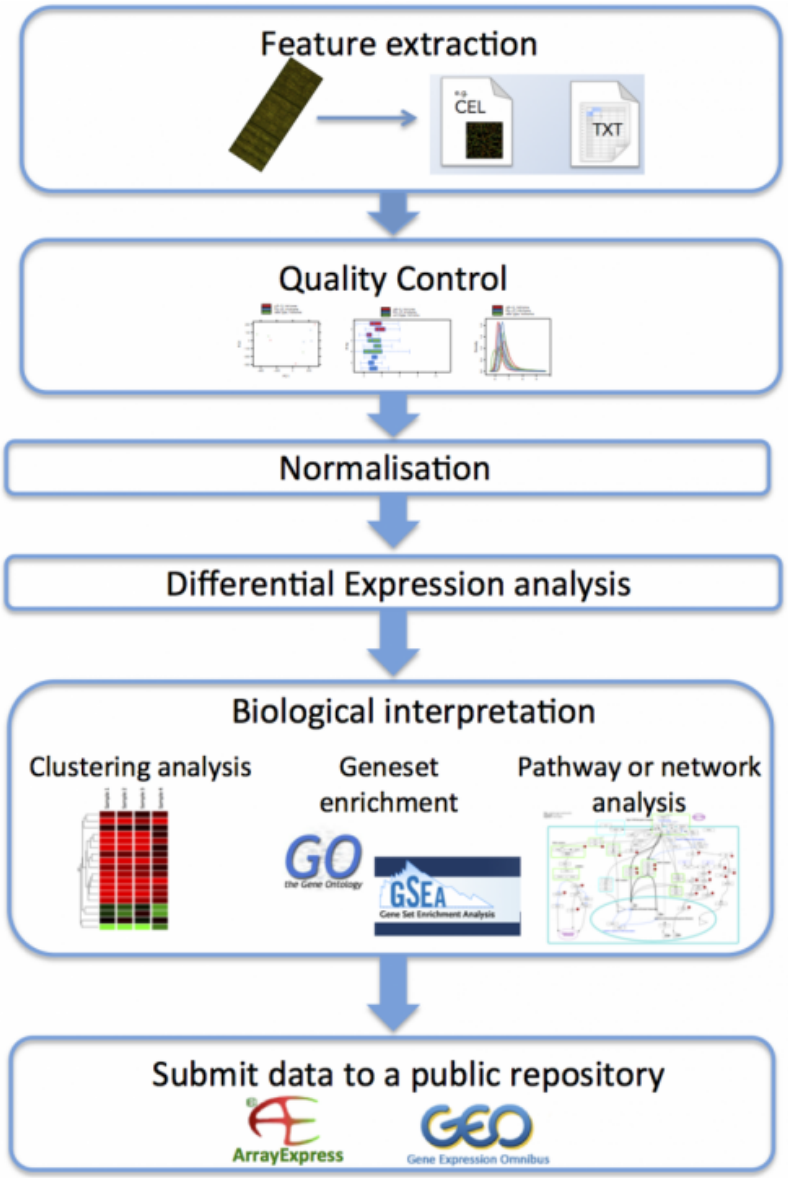
http://www.ebi.ac.uk/training/online/course/
functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

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.

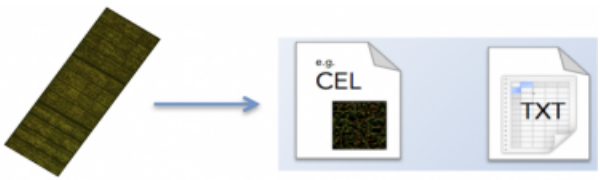
Both one and two colour microarrays can be used for this type of experiment.



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

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.)

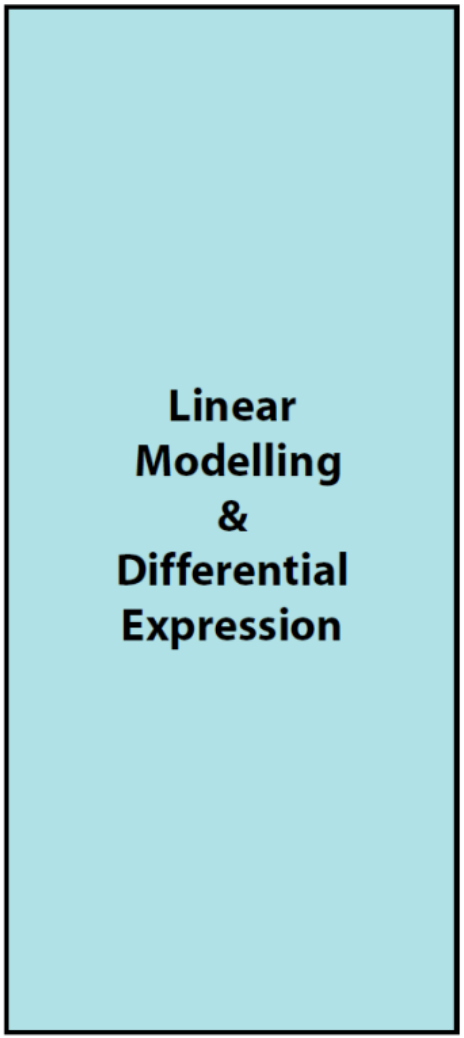
http://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

Limma Package

Step in Analysis	Function	Storage Class	
		1-colour	2-colour
Data Import	read.maimages / read.ilmn / read.idat	EListRaw	RGList
	readTargets / read.ilmn.targets		
	readGAL / readSpotTypes		
	controlStatus		
Preprocessing & Quality Assessment	backgroundCorrect / nec	EListRaw	RGList
	normalizeWithinArrays	EList	MAList
	normalizeBetweenArrays / neqc	EList	MAList
	voom / vooma / voomaByGroup	EList	
	plotMA		
	plotDensities		
	plotFB		
	imageplot		
	plotMDS		
	arrayWeights / voomWithQualityWeights		
	removeBatchEffect		

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

Limma Package



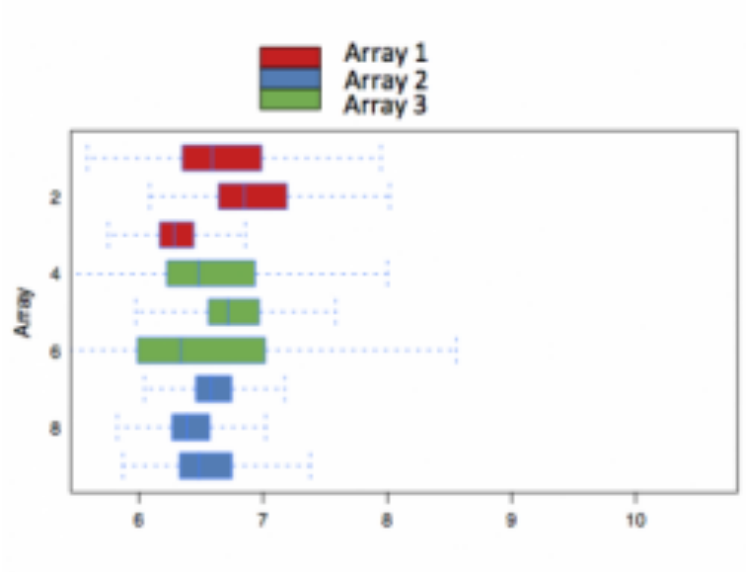
modelMatrix	
lmFit	MArrayLM
lmscFit	
avereps	
duplicateCorrelation	
makeContrasts	
contrasts.fit	
eBayes	
topTable	
treat	TestResults
topTreat	
decideTests	
write.fit	
propTrueNull	
genas	
volcanoplot	
heatdiagram / heatDiagram	
plotSA	
vennDiagram	

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

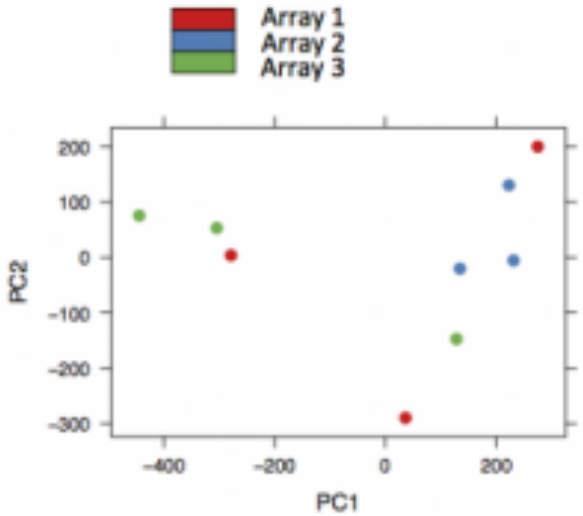
Quality control (QC)

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.

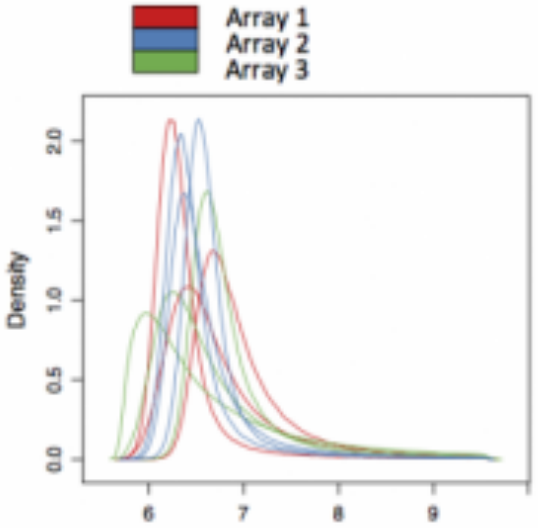
Data analysis software packages produce different sorts of diagnostic plots, e.g. of background signal, average intensity values and percentage of genes above background to help identify problematic arrays, reporters or samples.



Box plot



PCA



Density plot

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

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.

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

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.

Differential expression analysis: *t*-test

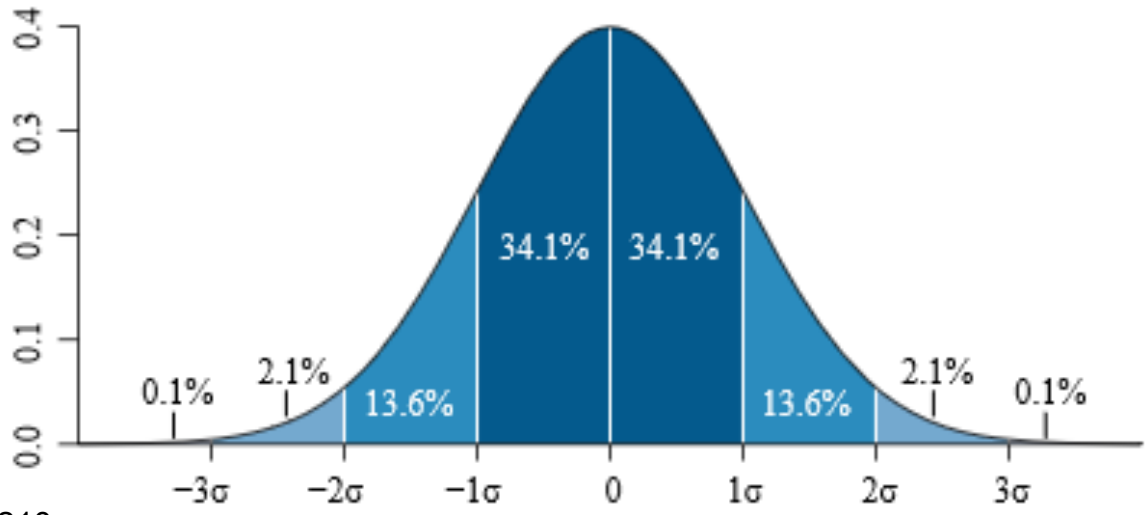
The *t* test is a simple, statistical method e.g. for detecting DE genes.

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)

Differential expression analysis: 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 noted above, 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.

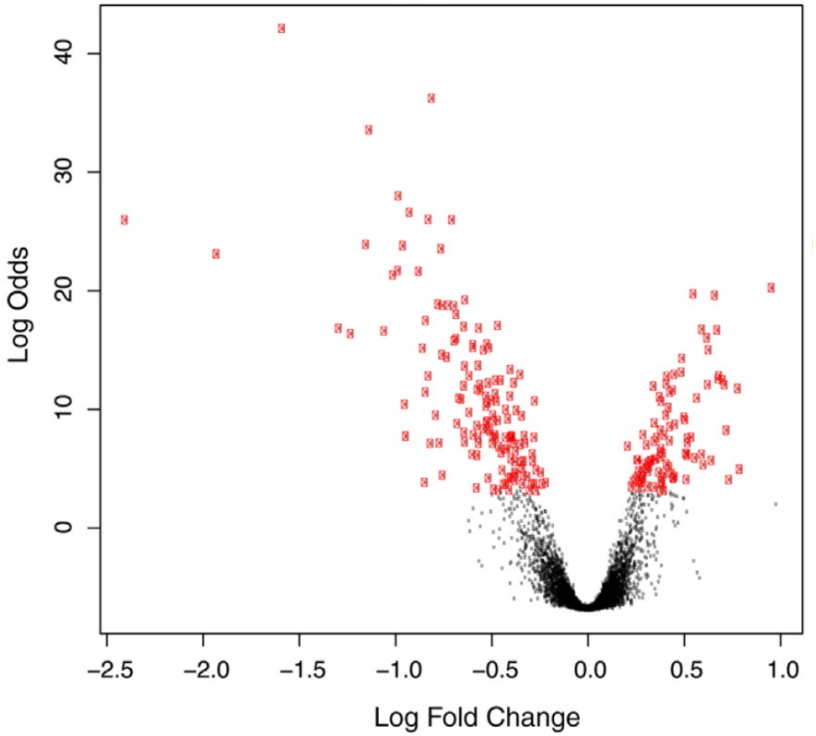
In the **'significance analysis of microarrays' (SAM)** version of the *t* test (known as the *S* test), a small positive constant *c* is added to the denominator of the gene-specific *t* test.

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 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

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**.

- 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

Example: Haemopedia

a Stem Cell Reports

Resource

OPEN ACCESS

Haemopedia: An Expression Atlas of Murine Hematopoietic Cells

Carolyn A. de Graaf,[1,5,*] Jarny Choi,[1,5] Tracey M. Baldwin,[1] Jessica E. Bolden,[1,5] Kirsten A. Fairfax,[1,5] Aaron J. Robinson,[1,5] Christine Biben,[1,5] Clare Morgan,[1,5] Kerry Ramsay,[1] Ashley P. Ng,[2,5] Maria Kauppi,[2,5] Elizabeth A. Kruse,[1,5] Tobias J. Sargeant,[1,5] Nick Seidenman,[1] Angela D'Amico,[3] Marthe C. D'Ombrain,[1,7] Erin C. Lucas,[1] Sandra Koernig,[7] Adriana Baz Morelli,[7] Michael J. Wilson,[7] Steven K. Dower,[7] Brenda Williams,[8,9] Shen Y. Heazlewood,[8,9] Yifang Hu,[4] Susan K. Nilsson,[8,9] Li Wu,[3,10] Gordon K. Smyth,[4,6] Warren S. Alexander,[2,5] and Douglas J. Hilton[1,5]

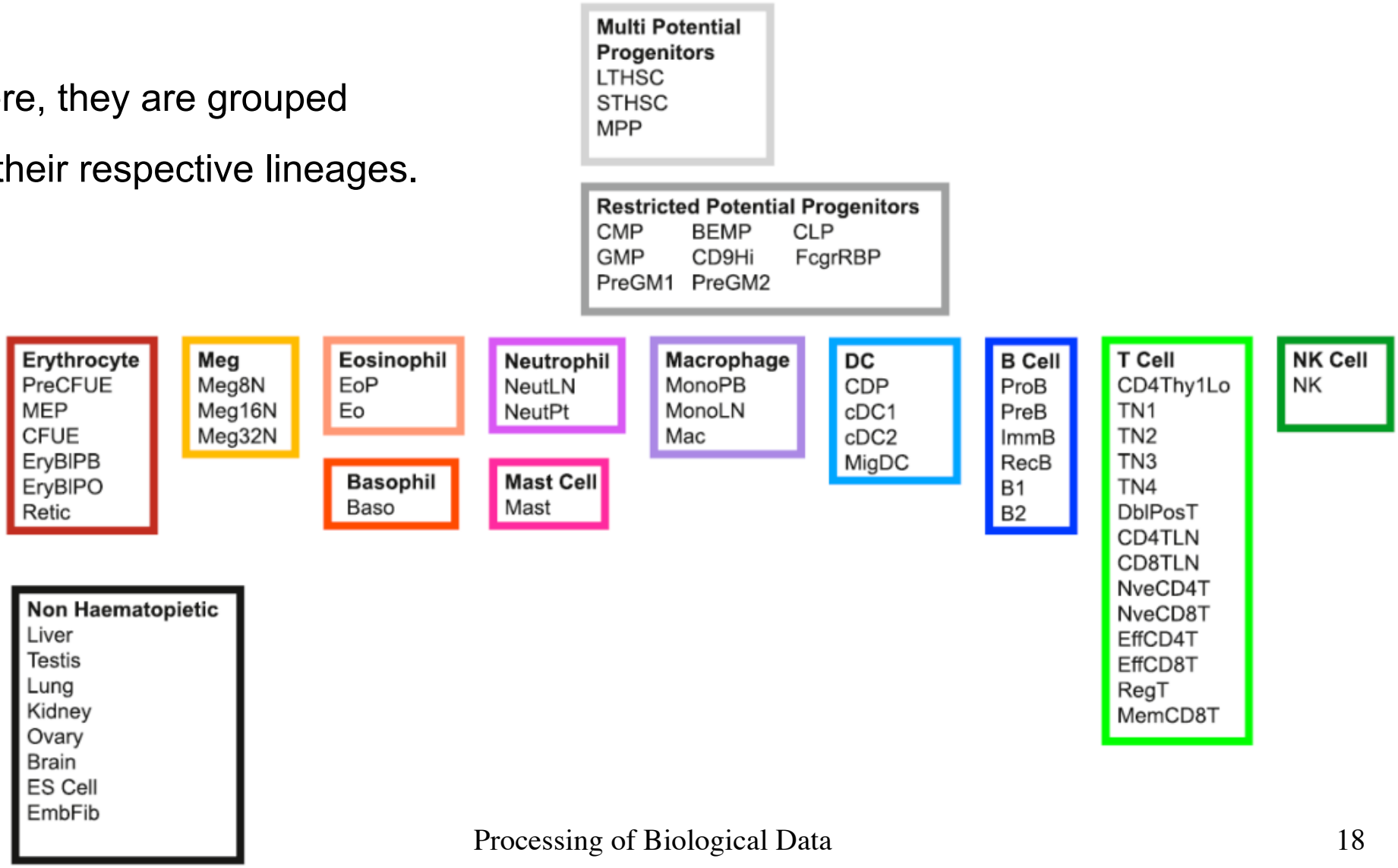
[1]Molecular Medicine Division
[2]Cancer and Haematology Division
[3]Molecular Immunology Division
The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia
[4]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3010, Australia
[5]Department of Medical Biology, University of Melbourne, Parkville, VIC 3010, Australia
[6]Department of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3052, Australia
[7]CSL Limited, Parkville, VIC 3052, Australia
[8]Biomedical Manufacturing, CSIRO Manufacturing, Clayton, VIC 3169, Australia
[9]Australian Regenerative Medicine Institute, Monash University, Clayton, VIC 3800, Australia
[10]Tsinghua University School of Medicine, Beijing 100084, China
*Correspondence: degraaf@wehi.edu.au
http://dx.doi.org/10.1016/j.stemcr.2016.07.007

Cells included in Haemopedia

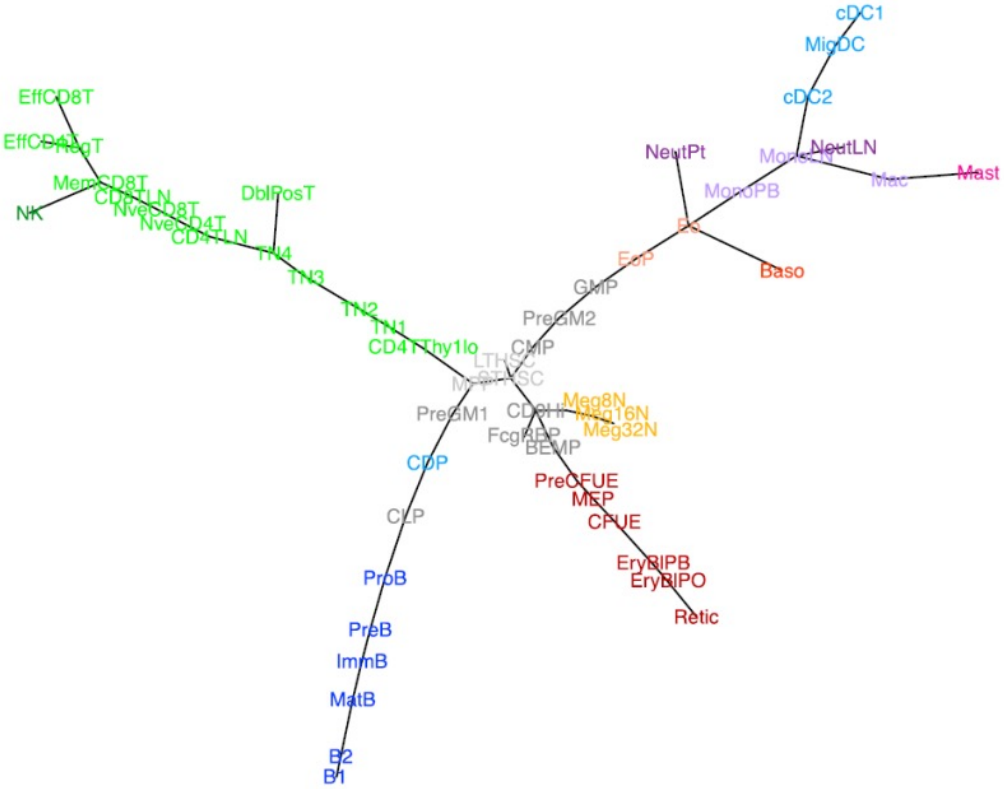
54 hematopoietic cell populations were purified by flow sorting from mouse and then analyzed by gene-expression profiling.

Here, they are grouped

in their respective lineages.



Relationship of Cells in Haemopedia



A total of 890 probes (719 genes) with SD > 2 on a $\log_2$ scale across all cell types were selected.

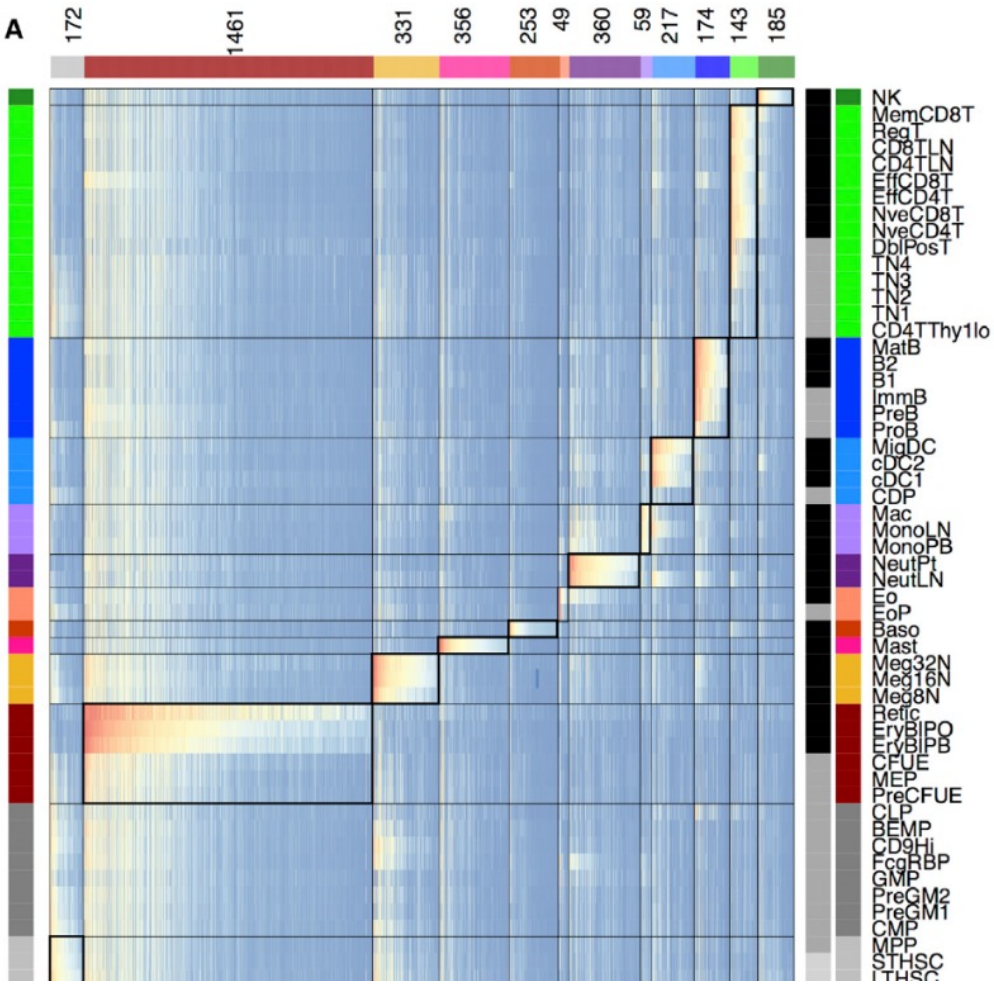
A **minimum spanning tree** based on Euclidean distance measurements was calculated using these probes. Lengths of branches reflect the distance between cell types.

Relationships of cells inferred by expression data recapitulate the progressive order of maturation; e.g.

- **megakaryocytes of increasing ploidy** (Meg8N, 16N–32N) or
- **T cell progenitor maturation** (CD4TThy1Lo [$Thy1^{lo}$ T cell progenitors], TN1, TN2, TN3 to TN4) or
- **B cell development** (ProB, PreB to ImmB).

Identifying Lineage-Specific Genes

Commitment, maturation, and activity of specific hematopoietic lineages are regulated by transcription factors and receptors that are expressed selectively



Lineage-specific genes: high expression in a single mature cell type and substantially lower expression in all other mature lineages

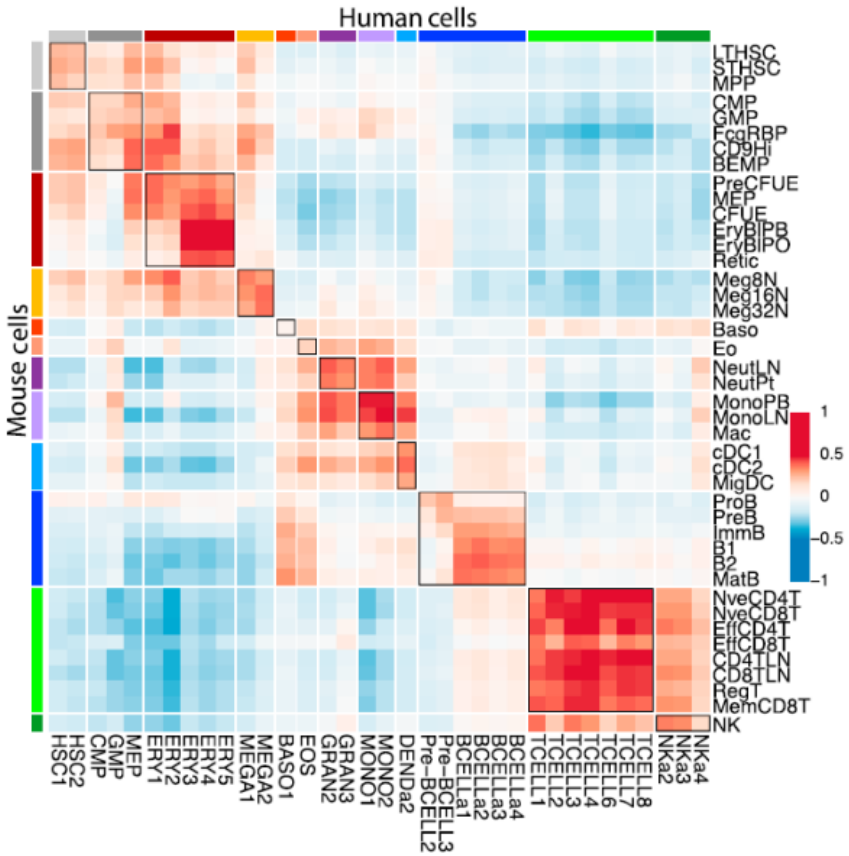
Heatmap is colored by the absolute expression value ($\log_2$) for each gene, **blue** is low, **yellow** intermediate, and **red** high expression.

Top line: number of genes specific for each lineage.

Mature cells are highlighted in black and progenitor cells in gray.

Heavily-lined boxes: expression of lineage signature genes in their associated lineage.

Expression in Mouse and Human Cells



Heatmap of correlations between mouse and human cell types after mean normalization of expression for ca. 9300 one-to-one orthologs between the species.

Genes with SD > 0.8 on a log2 scale were chosen, leaving 2,026 genes.

Heatmap scale is according to Pearson correlation of cell types, with no correlation (**dark blue**) through to highly correlated (**dark red**).

Lineages that are equivalent between species are highlighted by heavily lined boxes.

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.

Median absolute deviation (MAD) of a gene

MAD does not rely on the variance or standard deviation and thus it assumes no special statistical distribution of the data.

First the **raw median expression** for each gene *j* is calculated over all samples.

Then the **median absolute deviation** (MAD) of data points for this gene from its raw median is calculated as

$$MAD_i = median\left(\left|X_i - median_j\left(X_j\right)\right|\right)$$

Data points with maximum MAD are labeled as possible **outliers**

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 and then removes the point with the **maximum deviation** at each iteration.

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.

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

GESD

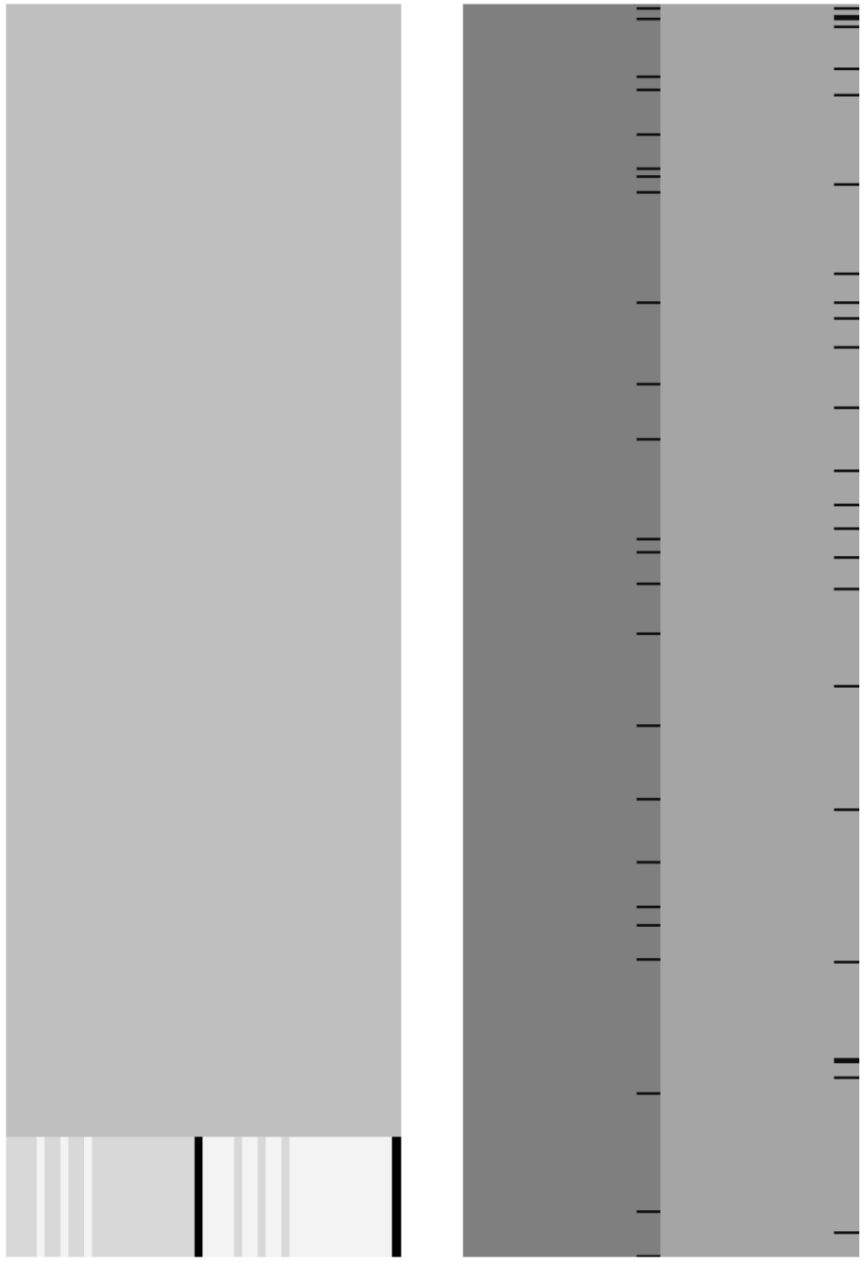
GESD and its predecessor ESD will **always mark** at least **one data point** as **outlier** even when there are in fact no outliers present.

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. The default in our work was half of the tested size.

Simulated expression data sets



Different gray levels represent different classes.

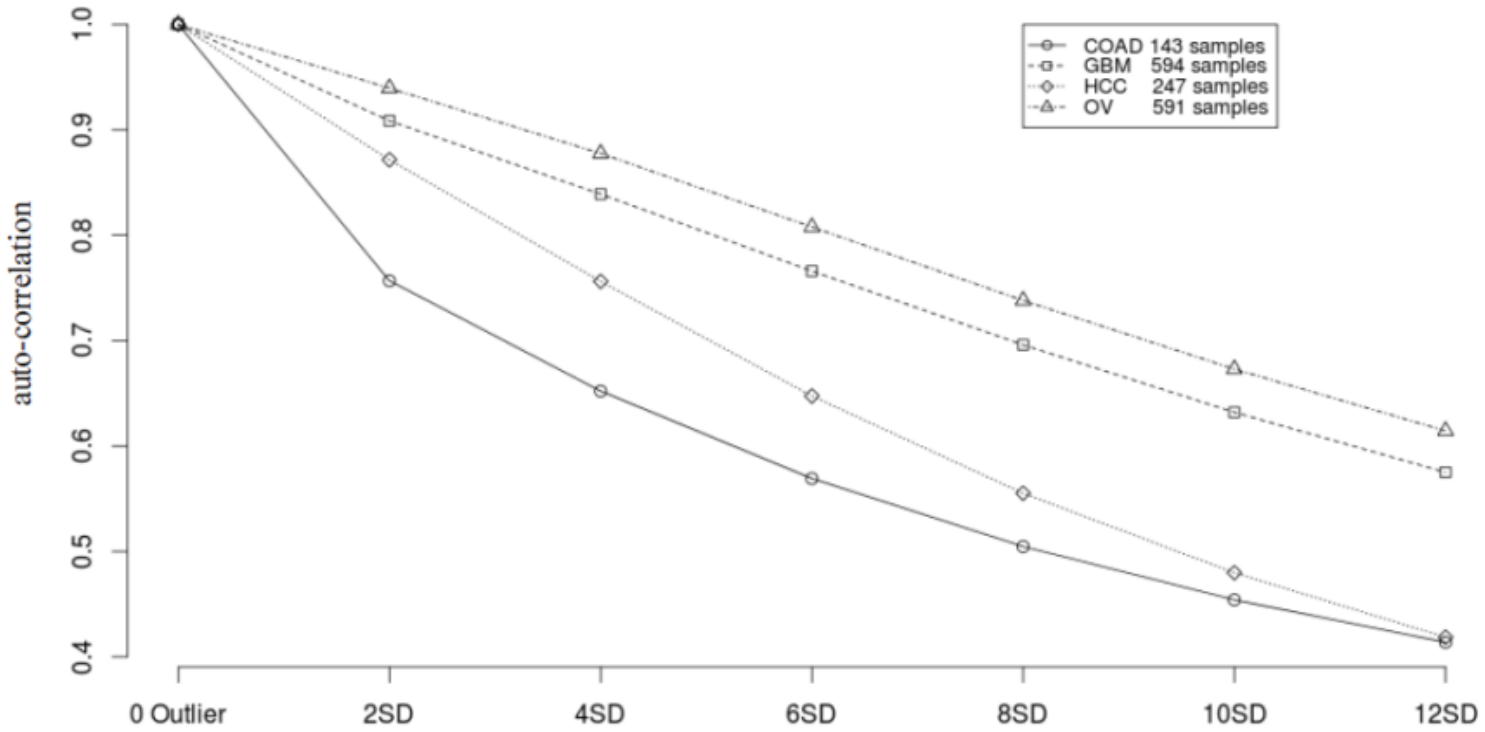
Outlier cases are in black.

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.

Effect of 2 outliers on auto-correlation of a gene



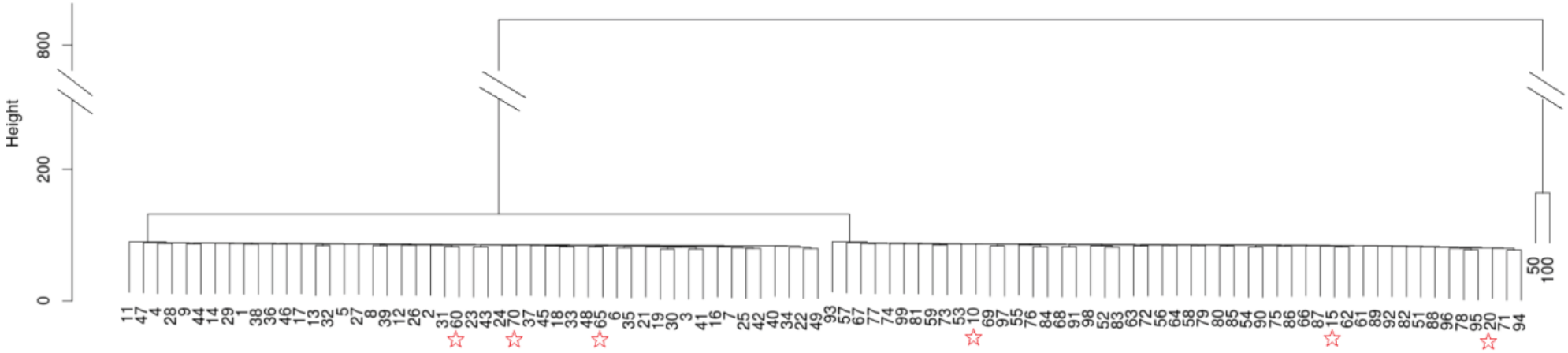
Effect of 2 introduced outlier points on co-expression analysis of a gene with itself

(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 correlation to 0.75.

Clustering dendogram

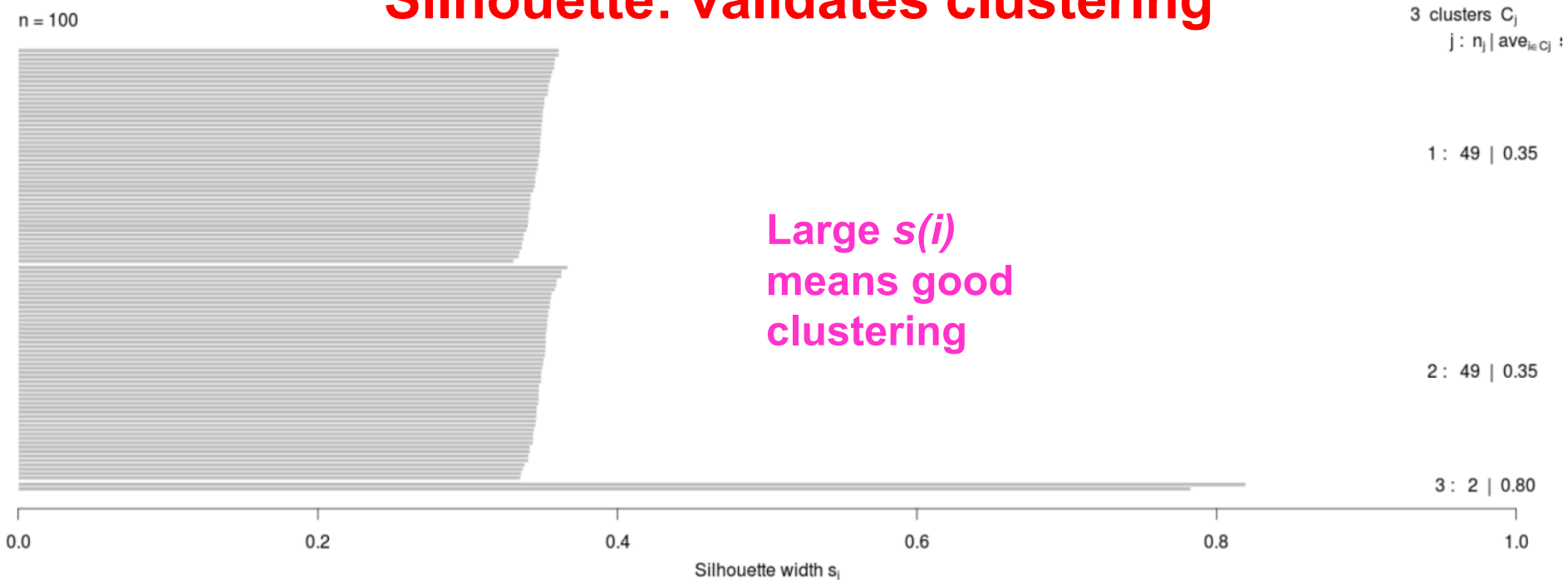


Clustering dendrogram of dataset of simulated expression.

Average Hierarchical Clustering bases 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.

Silhouette: validates clustering



Large *s(i)* means good 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

of detected synthetic outlier data points

	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.

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.

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

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

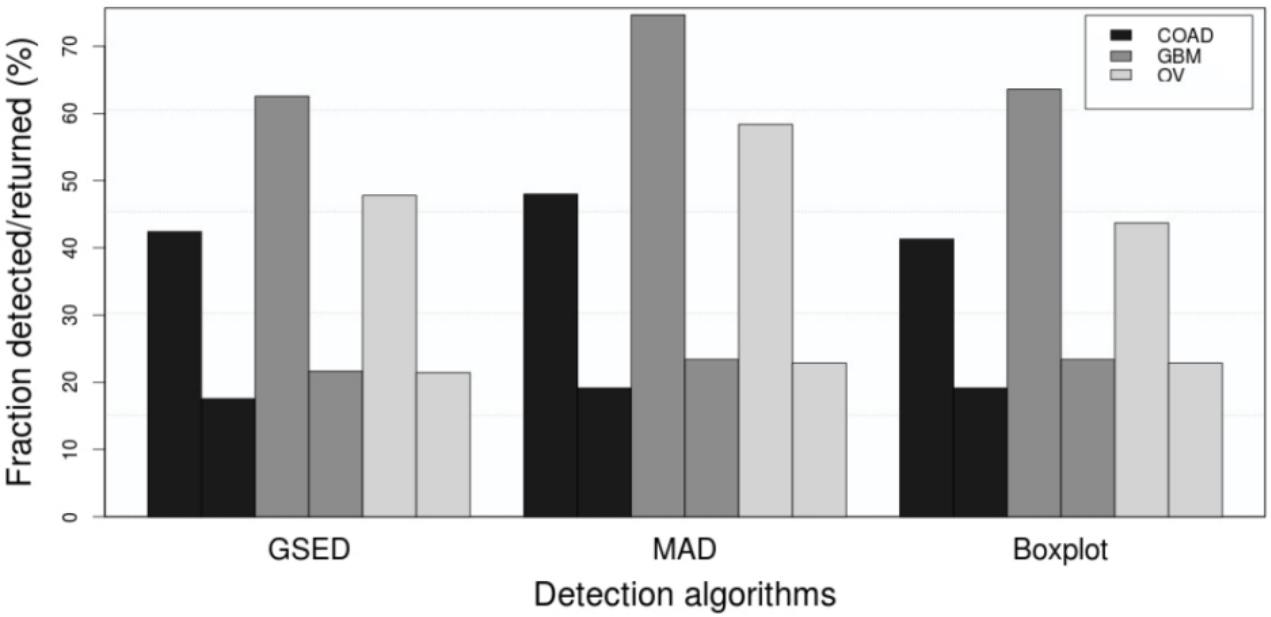
Clustering real data

Clusters found in TCGA colon expression dataset

Detected clusters in public colon cancer dataset from TCGA.

All 7 normal samples with barcode 11A were clustered together on the left side of the dendrogram away from tumor samples with barcode 01A.

Functionally relevant outliers



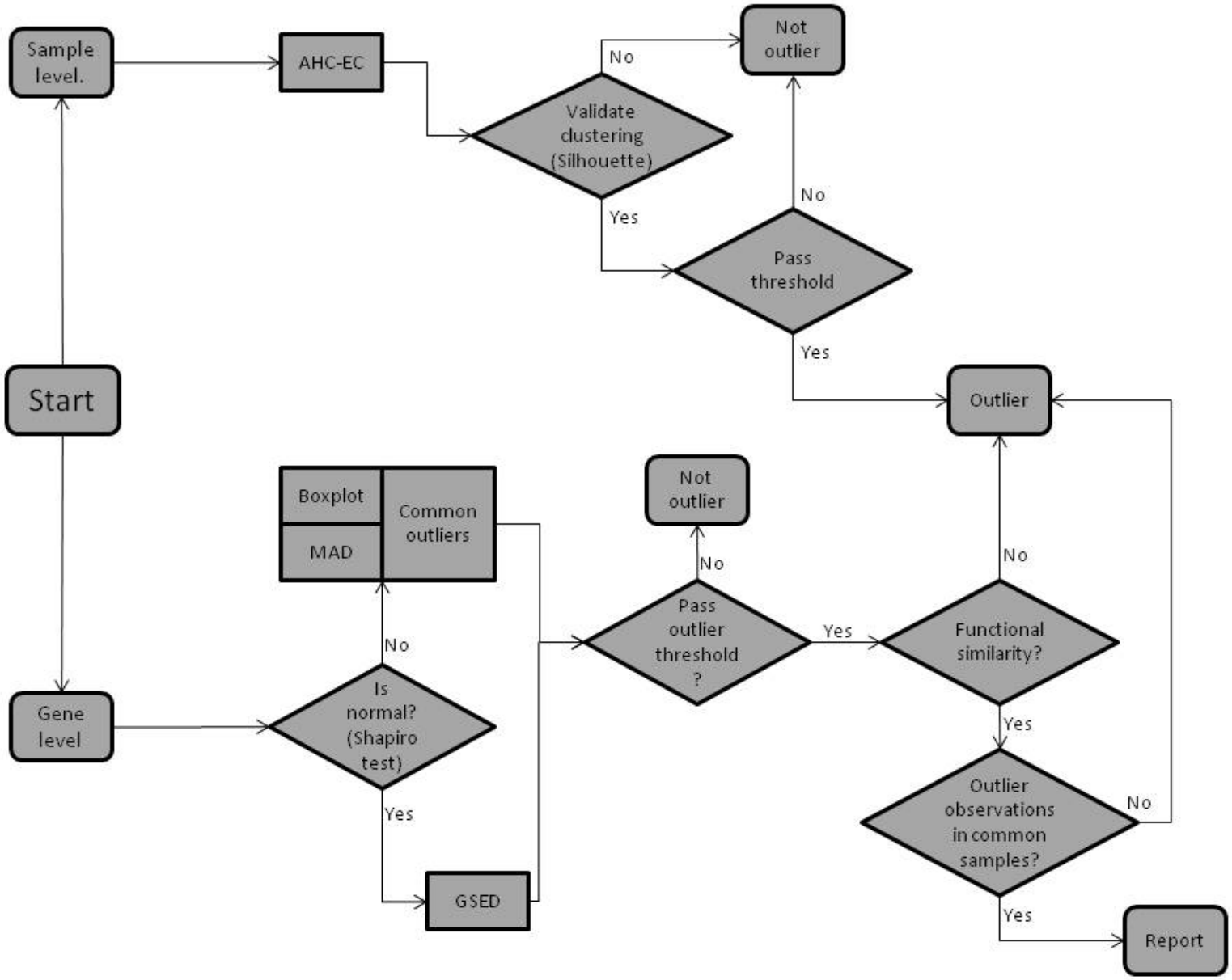
Idea: some outlier genes have functional similarity with other outlier genes in the same samplex and this may be functionally relevant.

Outlier detection statistics in TCGA methylation datasets.

Percentage of detected and returned outliers - due to functional similarity (from GOSemSim package, see V8) and common positions - in the TCGA methylation datasets COAD, GBM and OV.

The left column in each group refers to the fraction of detected and the right column refers to the fraction of returned outliers.

Our workflow



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, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK

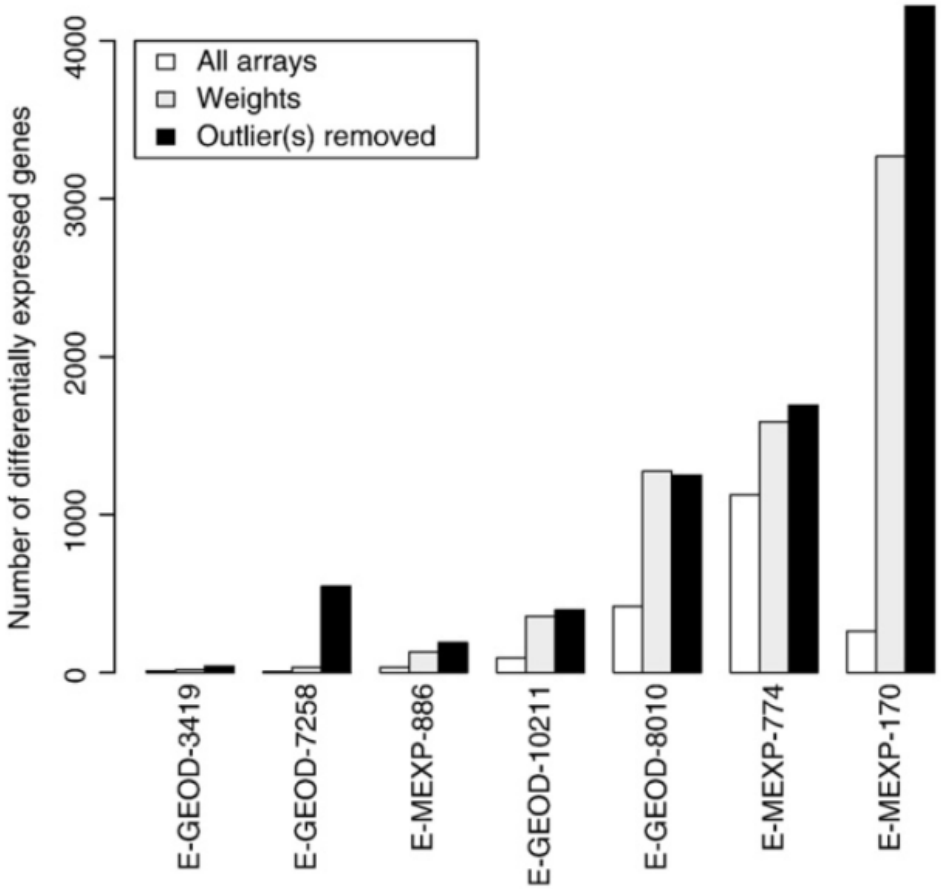
These authors compared four strategies of data analysis :

- 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

Number of DE genes

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



Kauffman, Huber (2010) Genomics 95, 138

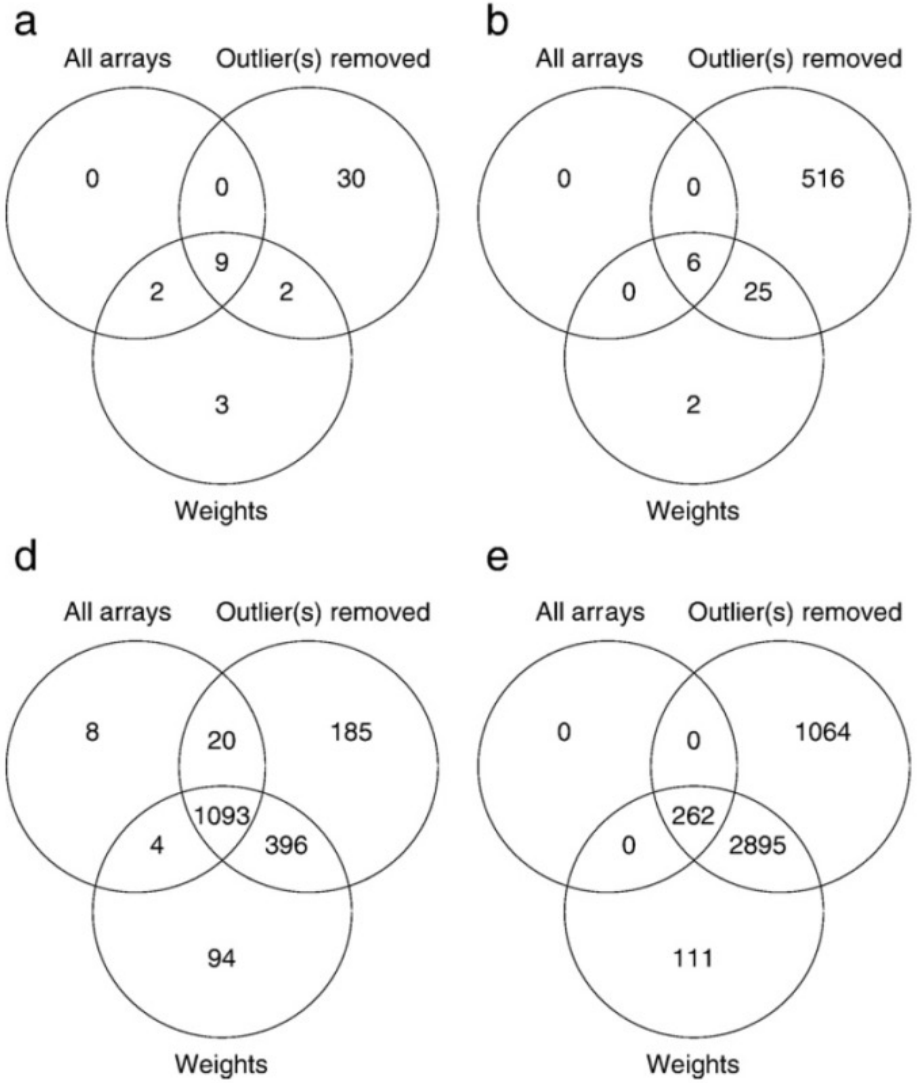
Number of differentially expressed genes identified:

- 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 confound-ding effect of experiment date! This explains high # of DE genes.

Effect of Outlier removal on DE genes



Kauffman, Huber (2010) Genomics 95, 138

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,

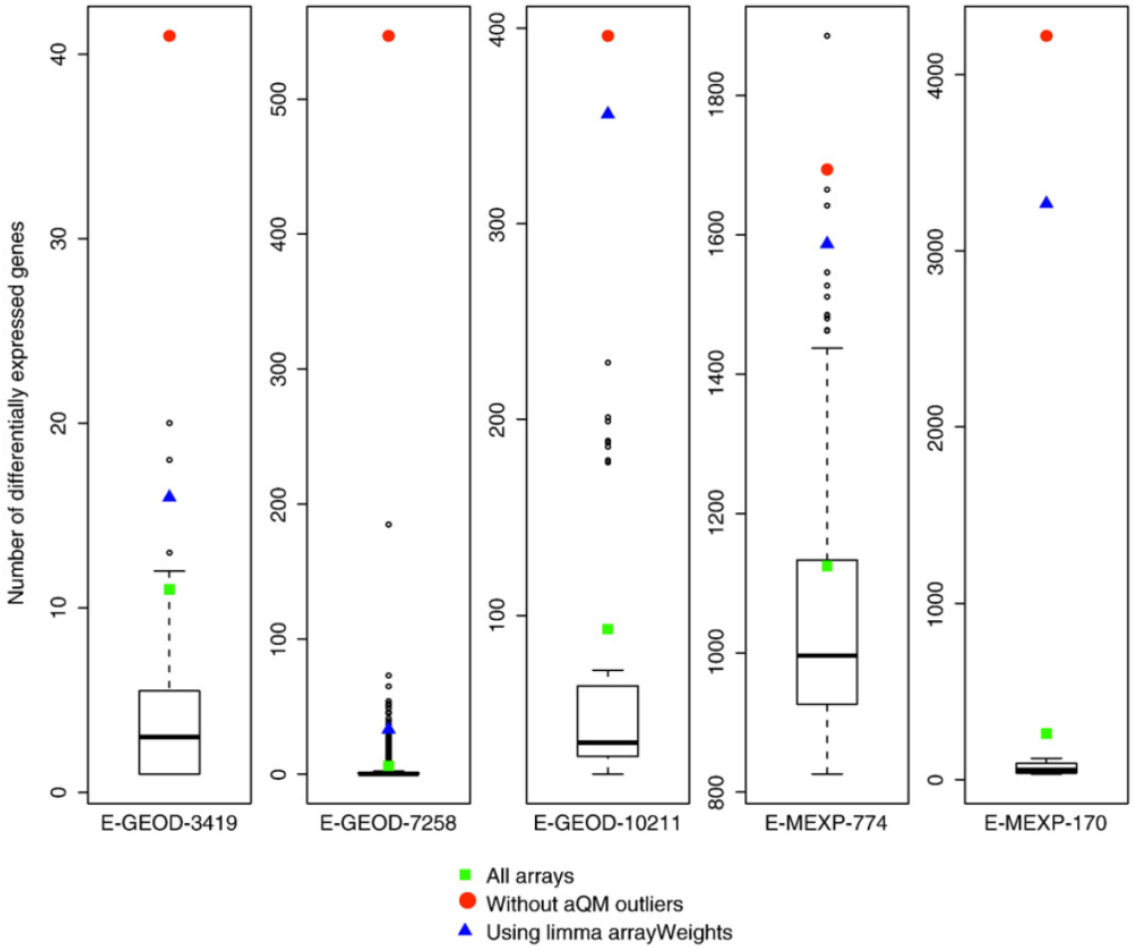
(b) E-GEOD-7258,

(c) E-GEOD-10211,

(d) E-MEXP-774,

(e) E-MEXP-170.

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

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

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 (it is 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

DE detection in 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.

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

DE detection in RNAseq data

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

To address this over-dispersion problem, methods such as edgeR and DESeq use the related **negative binomial distribution** (NB)
where the relation between the variance *v* and mean μ is defined as

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

where α is the **dispersion factor**.

Estimation of this factor is one of the fundamental differences between the edgeR and DESeq packages.

NB distribution:

$$f(k; r, p) \equiv \Pr(X = k) = \binom{k+r-1}{k} p^k (1-p)^r \quad \text{for } k = 0, 1, 2, \ldots$$

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

Reference data

Samples from **group A** : Strategene Universal Human Reference RNA (UHRR), which is composed of 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 concentrations of 4 subgroups of the synthetic spike-ins such that the log expression change is predefined and can be used to benchmark DE performance.

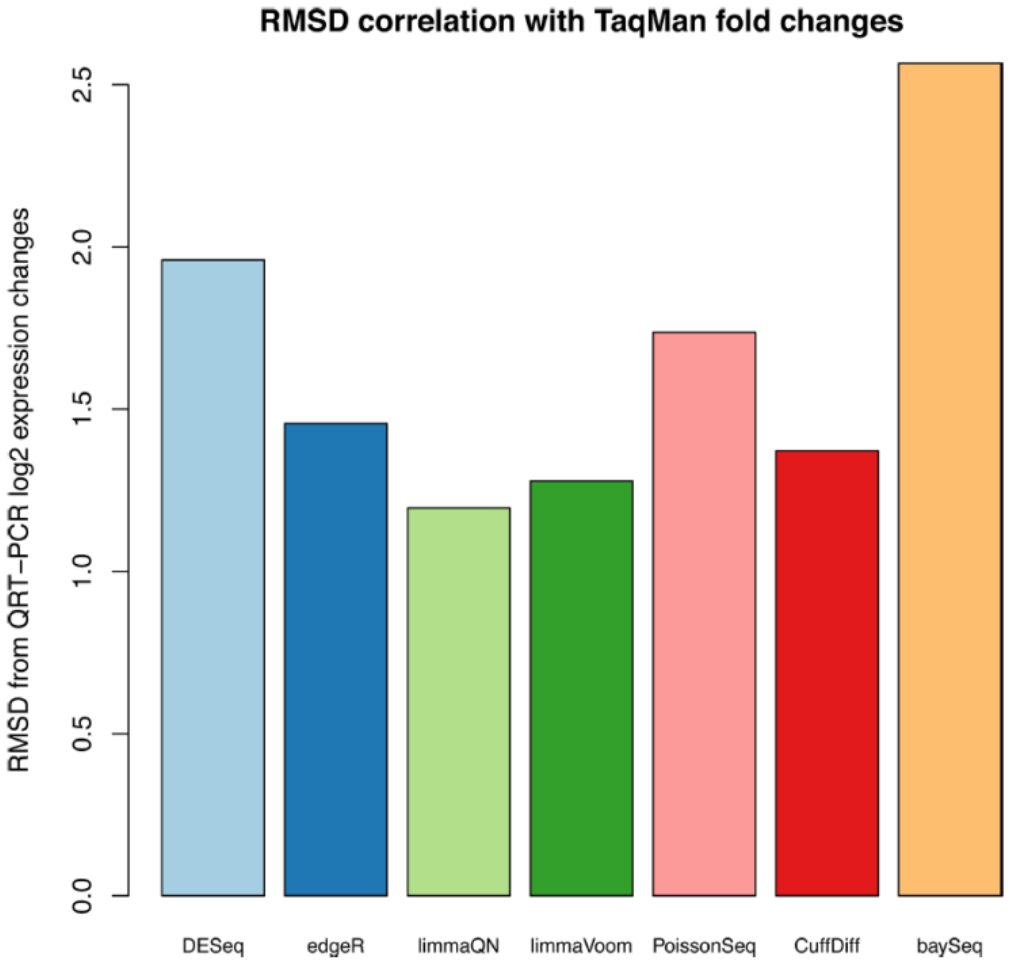
Rapaport et al. (2013) Genome Biol. 14: R95

Comparison against reference data

RMSD correlation between qRT-PCR and RNA-seq $\log_2$ expression changes computed by each method.

Overall, there is **good concordance** between $\log_2$ values derived from the DE methods and the experimental values derived from qRT-PCR measures.

Upper quartile normalization (baySeq) is least correlated (highest RMSD) with qRT-PCR values.



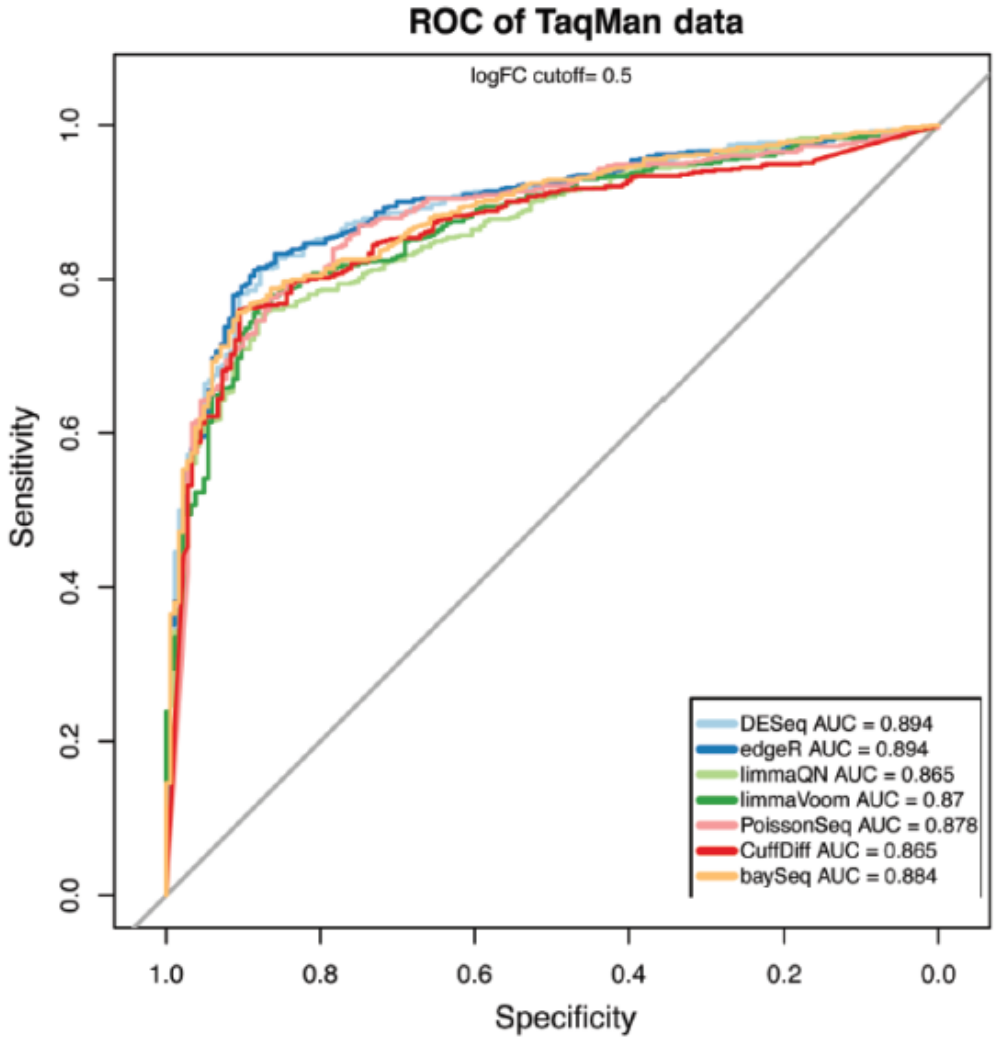
Rapaport et al. (2013) Genome Biol. 14: R95

Performance for DE detection

Differential expression analysis using qRT-PCR validated gene set.

ROC analysis was performed using a qRT-PCR $\log_2$ expression change threshold of 0.5.

The results show a slight advantage for **DESeq** and **edgeR** in detection accuracy.



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

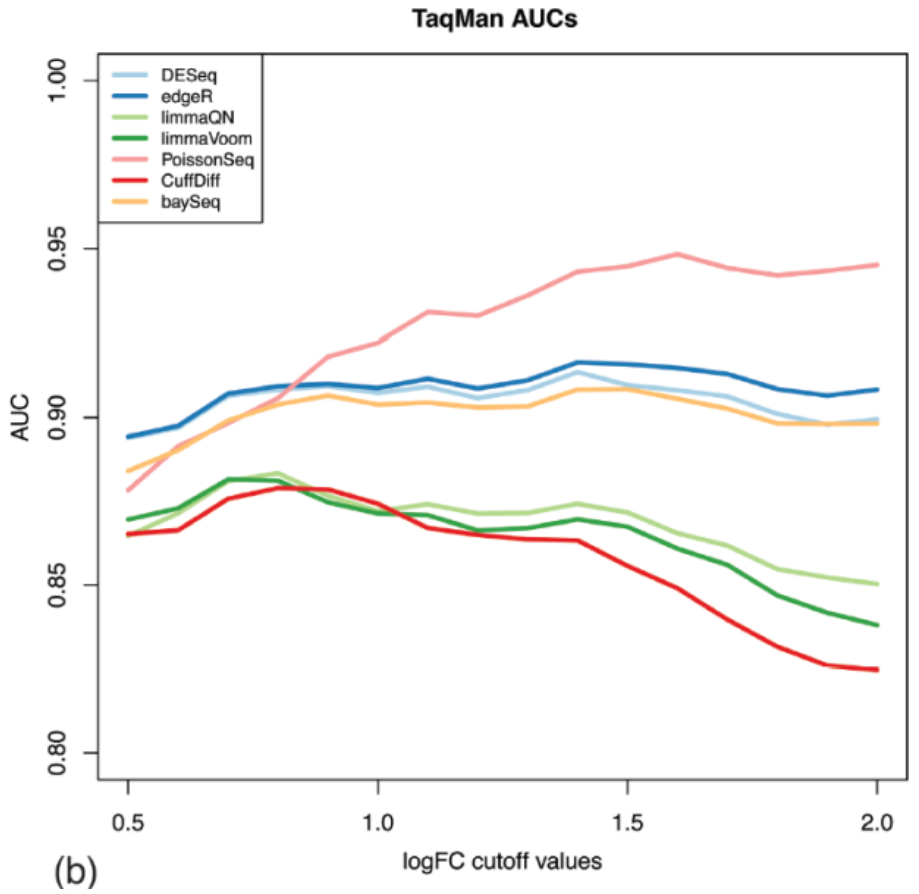
Performance for different thresholds

Increasing $\log_2$ expression ratios represent a more stringent cutoff for differential expression.

→ one would expect a better performance of the DE methods.

Indeed, the performance of **PoissonSeq** increases, whereas that of the **Cuffdiff** and **limma methods** gradually reduce.

AUC, area under the curve.



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

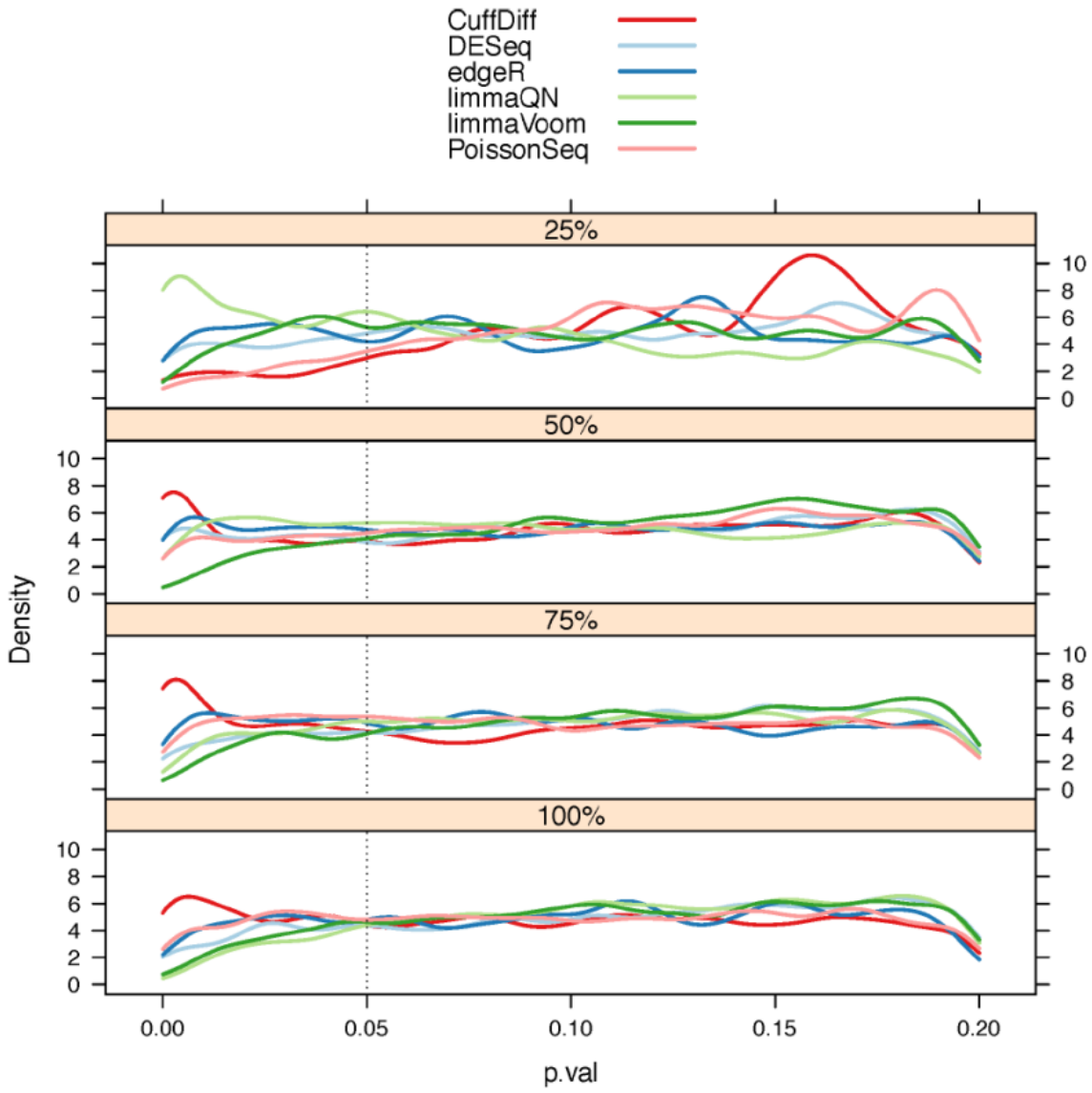
Intra-condition comparisons

Intra-condition comparisons using the SEQC technical replicate samples from each condition.

No DE genes are expected in these comparisons.

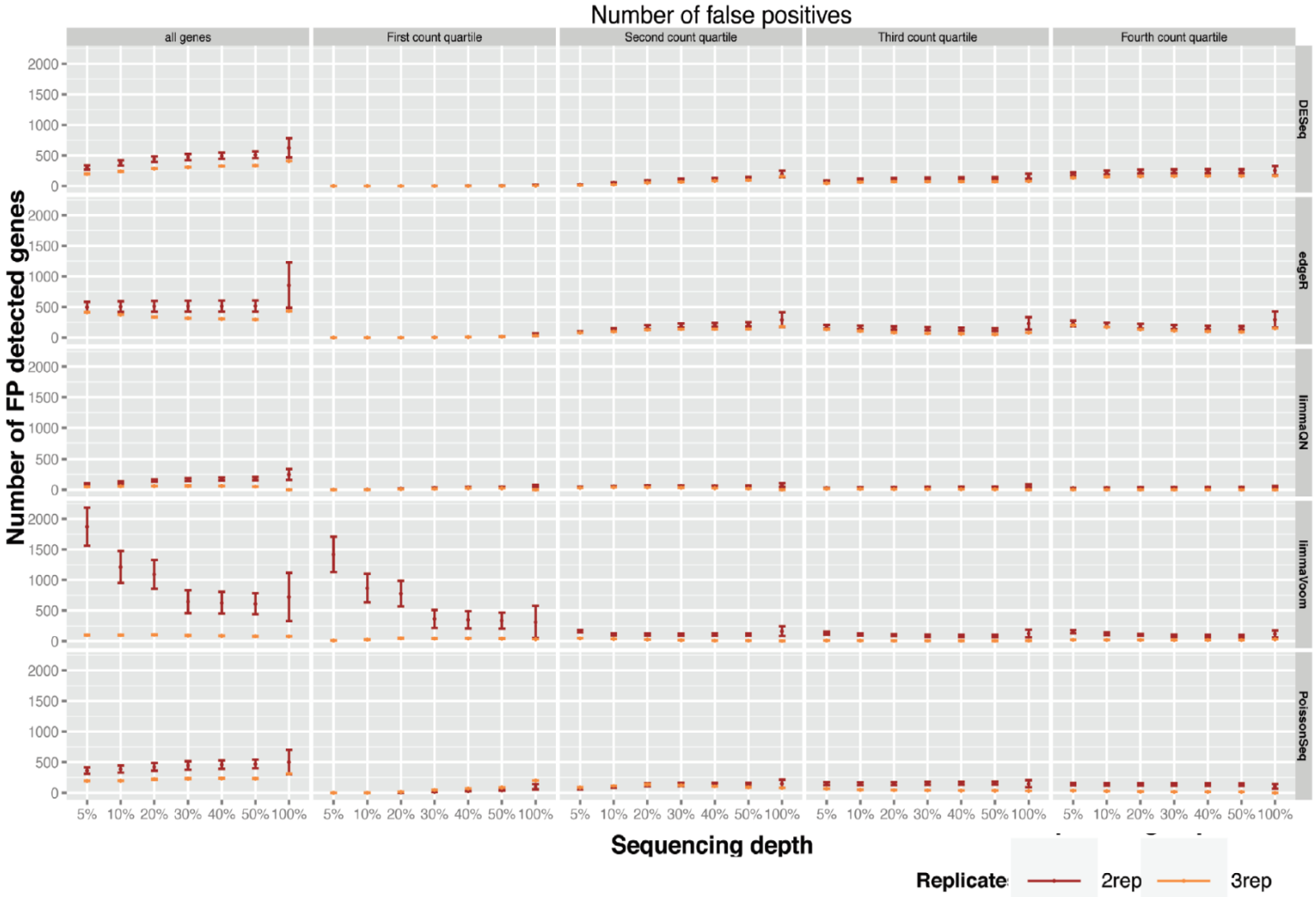
The distribution of P values is expected to be uniform since they are derived from the null model.

Indeed, we found that the P values for all methods were largely uniform



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

As expected, all methods had a smaller number of FPs with increasing number of replications and increased sequencing depths.



FP calls among the lowest 25% of expressed genes increased with sequencing depth and number of replicates in contrast to the higher expression quartile where the FP rate reduces when more data is provided. However, the total number of FPs is lowest in the bottom 25% expression indicating that all methods are **conservative** when predicting DE at low expression ranges.

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

(b)

Sensitivity rates defined as the fraction of true set genes **improves significantly** with the sequencing depth and number of replicates.

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

Comparison of methods

Table 2 Comparison of methods.

Evaluation	Cuffdiff	DESeq	edgeR	limmaVoom	PoissonSeq	baySeq
Normalization and clustering	All methods performed equally well					
DE detection accuracy measured by AUC at increasing qRT-PCR cutoff	Decreasing	Consistent	Consistent	Decreasing	Increases up to log expression change ≤ 2.0	Consistent
Null model type I error	High number of FPs	Low number of FPs	Low number of FPs	Low Number of FPs	Low number of FPs	Low number of FPs
Signal-to-noise vs *P* value correlation for genes detected in one condition	Poor	Poor	Poor	Good	Moderate	Good
Support for multi-factored experiments	No	Yes	Yes	Yes	No	No
Support DE detection without replicated samples	Yes	Yes	Yes	No	Yes	No
Detection of differential isoforms	Yes	No	No	No	No	No
Runtime for experiments with three to five replicates on a 12 dual-core 3.33 GHz, 100 G RAM server	Hours	Minutes	Minutes	Minutes	Seconds	Hours

AUC, area under curve; DE, differential expression; FP, false positive.

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

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.

Robust outlier detection methods for RNA-seq data should yield better **performance** expected for higher number of replicates + sequencing depth.